# Forecasting the space heating demand of Dutch households

Master thesis Econometrics & Management Science
Business Analytics & Quantitative Marketing

August 5, 2020


Erasmus School of Economics
Erasmus University Rotterdam

*Author*
A.P. (Toon) Jansen

*Student ID*
475941

*Supervisor*
dr. W. Wang

*Second assessor*
prof. dr. P.H.B.F. Franses

*External supervisor*
T. van de Sande MSc

## Abstract

The world is facing a transformation of the energy sector. A main challenge in this transformation is heating the built environment. District heating networks are a key solution in this challenge. As a consequence, energy companies investing in heat networks are wondering what the future heat demand will be, given a changing climate and ongoing energy efficiency gains in the building environment. In this research, the future heat demand of Dutch households is investigated, with 2050 as time horizon. Using yearly space heating demand data at household level, average household demand is forecast applying forecast combination with three econometric methods. Results for the random effects model, the functional coefficient model and the recurrent neural network show that degree days are the most important predictor for space heating demand and that, as a result, rising temperatures will strongly affect this demand. In addition, the isolation quality, floor surface, building year and house type affect the space heating demand for households. The space heating demand is expected to decrease with 0.8% on a yearly basis, resulting in an expected space heating demand of 22.4 GJ in 2050. The 90% confidence interval ranges from 17.2 to 29.7 GJ in 2050. To improve the forecasting accuracy, most relevant would be to find a way to model occupant behaviour. In addition, forecasts could likely be improved using a higher temporal granularity.



**Keywords:** Space heating demand, District heating networks, Random Effect Model, Functional Coefficient Model, Recurrent Neural Network

## Preface

This work is written as a master thesis for the study programme Econometrics & Management Science at the Erasmus University Rotterdam. Having the wish to contribute to the energy transition, I decided to combine my thesis with an internship at Eneco, an energy company. I investigated the future heat demand of Dutch households, supporting Eneco in their decision making about district heating networks. This thesis is the result. I would like to thank Wendun Wang for his useful feedback, and Thomas van de Sande for the many chats and good advise on how to present my results and how to make my research valuable in practice.

I wish you a pleasant read!

Toon Jansen

The Hague, August 5, 2020

# Contents

## List of abbreviations

| Abbrevation | Meaning |
|---|---|
| AIC | Akaike Information Criterion |
| ANN | Artificial Neural Network |
| BIC | Bayesian Information Criterion |
| CBS | Central Bureau for Statistics |
| FCM | Functional Coefficient Model |
| FE | Fixed Effects |
| GH | KNMI climate scenario with moderate worldwide temperature increase and a high impact of air stream pattern changes in the Netherlands |
| GJ | Gigajoule |
| GH | KNMI climate scenario with moderate worldwide temperature increase and a low impact of air stream pattern changes in the Netherlands |
| IPCC | International Panel on Climate Change |
| LSTM | Long short-term Memory (cell) |
| KNMI | Royal Dutch Weather Institute |
| MAR | Missing At Random |
| MAPE | Mean Absolute Percentage Error |
| MAPPE | Mean Absolute Percentage Prediction Error |
| MCAR | Missing Completely At Random |
| MCD | Minimum Covariance Determinant |
| MNAR | Missing Not At Random |
| MSE | Mean Squared Error |
| MSPE | Mean Squared Prediction Error |
| NARX | Non-linear Auto-regressive exogenous |
| OLS | Ordinary Least Squares |
| RE | Random Effects |
| RNN | Recurrent Neural Network |
| SARIMA | seasonal auto-regressive integrated moving average |
| SIM | Single-index model |
| SVM | Support Vector Machine |
| WH | KNMI climate scenario with strong worldwide temperature increase and a high impact of air stream pattern changes in the Netherlands |
| WL | KNMI climate scenario with strong worldwide temperature increase and a low impact of air stream pattern changes in the Netherlands |

# 1 Introduction and Research Questions

Climate change is considered one of the world's biggest problems and challenges, especially in the scientific community. In 2019, more than 10.000 scientists warned the world of a climate emergency (Ripple, Wolf, Newsome, Barnard, & Moomaw, 2019). Climate change is largely, but not uniquely, a problem related to energy. To mitigate climate change, the world needs to decarbonise its energy production and to increase its energy efficiency. Buildings are an important part of this challenge, as they account for a significant part of energy use and $CO_2$-emission. In Europe, buildings make up 40% of energy use and 36% of $CO_2$-emissions (European Parliament and Council, 2010).

In the Netherlands in particular, decarbonising the housing sector is difficult. The Netherlands is considered addicted to gas, since the discovery of a one the largest gas reserves worldwide in 1959 (De Groene Amsterdammer, 2018; Energeia, 2020). This has led to an extensive gas infrastructure, and gas has been supplied to almost all households for heating and cooking ever since. However, as a consequence of climate change and limitations in the Dutch gas supply, the Dutch government wants households to quit using gas for heating and cooking (De Nederlandse Rijksoverheid, 2020). Delivering heat via district heating networks is one the most prominent solutions to replace gas. More in general, district heating networks are considered "essential key elements of future sustainable energy systems" (Koschwitz, Spinnräker, Frisch, & van Treeck, 2020). At this moment however, heat infrastructure only exists in some of the larger cities and its surburbs in the Netherlands.

As a consequence, energy companies are starting to develop district heating networks and they want to gain insight in future heat demand. They foresee that the heat demand will be affected negatively by climate change and insulation improvements. Climate change causes higher temperatures and could thus cause less heat demand (Andrić et al., 2016). Better insulation is an important pillar of energy efficiency in the built environment (Wilkinson, Smith, Beevers, Tonne, & Oreszczyn, 2007). Improved building insulation levels could decrease heating demand even more than climate change (Andrić et al., 2016).

"The heat demand in a district heating system generally originates from space heating and tap water heating. While space heating is primarily weather dependent, the tap water usage is related to social behaviour." (Frederiksen & Werner, 2013). Tap water heating includes the heating of all water in households, for showering, doing dishes, washing etc. The social component of hot tap water demand causes mostly variations within days (e.g. when do people shower) and weeks (e.g. when do people wash clothes), and not between years (Fang & Lahdelma, 2016). Space heating is done via radiators and is important for people to

live in a comfortable temperature. A third component of heat production are heat losses during transport. These are, however, usually not included in the heat demand, as the amount of heat purchased is measured at the client. As a consequence of all the aforementioned, it is expected that only the demand for space heating is influenced by climate change and improvements in insulation. The absence of a relation between hot tap water demand and outside temperature is presented in Appendix A.2. In addition, the variations caused by social behaviour do not affect yearly demand, and are thus not relevant for long term predictions.

The temperature and insulation levels are thus expected to influence the future space heating demand in particular. However, it is either uncommon or difficult to link these factors directly to space heating demand. When the outside temperature reaches above the desired inside temperature, space heating demand is likely to disappear. Hence, space heating demand is usually more directly related to degree days. Degree days, simply said, count the number of degrees below a certain threshold temperature. The idea is that people start heating their homes when the temperature drops below this threshold. Insulation quality, on the other hand, is an abstract concept that is very difficult to measure directly. A recently established general measure of energy efficiency is the energy label, which ranges from A (very energy efficient) to G (very energy inefficient) (Rijksoverheid, 2020b).

In this paper, the future demand for space heating of Dutch households is investigated, with the year 2050 as time horizon. Multiple econometric models are compared to find reliable estimates. Forecasting space heating demand of various types of Dutch households is the main goal of the study. In order to do this, a relationship between degree days, a building's energy label and other building characteristics on the one hand and space heating demand on the other hand needs to be established. This part is the main focus of this research. Reliable data or scenarios from other sources for degree days and energy labels in the period 2020 to 2050 need to be found in order to use this as input for the forecasting. Finally, a side goal of this research is to find an appropriate econometric method to model the relation between the aforementioned variables. These goals of the study lead to the three research questions below.

1. What is the effect of degree days, a building's energy label and other building characteristics on space heating demand?

2. What is the expected space heating demand for a household from 2020 until 2050 and what is the variability of this prediction?

3. What is the most appropriate econometric method to model the relationship between degree days, a building's energy label and other building characteristics on the one hand,

and space heating demand on the other hand?

Section 2 gives a review of existing literature regarding the research questions. In section 3 the available data is discussed and in section 4 the methodology to analyse these data is presented. Section 5 shows the results obtained. In section 6 this research is concluded and discussed.

# 2 Literature review

In the literature review, several topics pass in review. First of all, space heating demand modelling is dealt with extensively. Secondly, temperature and degree days are discussed. Thirdly, insulation and energy labels are looked into. In the last section, literature on the forecasting of space heating demand is presented.

## 2.1 Space heating demand

In modelling space heating demand there is a distinction to make between engineering or physical models on the one hand and data-driven models on the other hand (Amasyali & El-Gohary, 2018). Physical models rely on thermodynamic rules and require very detailed physical information, which is often not available for many buildings. As a consequence, frequently a few example buildings are analysed into detail. For this study, a data-driven model is used, as the idea is to make predictions for a variety of building types, not only for a specific type of building.

The data-driven models to predict energy consumption have become increasingly popular in research during the recent years (Amasyali & El-Gohary, 2018). These approaches require less detailed information, but are considered black-box models. Data-driven models to predict energy consumption have been reviewed to gain insight in the research gaps by Amasyali and El-Gohary (2018). They found that only 19% of research investigated residential buildings, and only 12% had the temporal granularity of a year. It was, however, argued that despite their challenges, both long-term energy consumption prediction models and residential consumption prediction models are needed (Amasyali & El-Gohary, 2018). Wei et al. (2018) found that most researches investigating residential buildings are at low granularity, i.e. energy consumption is estimated on a regional level. Therefore, this research addresses a knowledge gap in scientific research by investigating long-term residential space heating consumption at a high spatial granularity. Furthermore, it has the potential to contribute to developing district heating networks, and thus to the (Dutch) energy transition. In the next sections different data-driven methods are discussed. The section thereafter deals with different predictors.

### 2.1.1 Methods

In forecasting space heating demand methods, a distinction has to be made between machine learning methods and statistical methods. Machine learning methods are used more often for short term forecasting, while statistical methods are used more often for long-term fore-

casting. For both types of methods there is a vast amount of more specific variants (Zhao & Magoulès, 2012). Machine learning methods can be very complex but are more flexible and hence often better at predicting than statistical regression methods (Wei et al., 2018). However, for understanding the structure of the relations, or for interpretability, the statistical methods usually perform better (Mastrucci, Baume, Stazi, & Leopold, 2014). This section first discusses statistical models, then machine learning models and finally two other methods that are somewhat more complex than the statistical methods, but less complex than the machine learning methods.

### Statistical methods

Fang and Lahdelma (2016) compared a seasonal auto-regressive integrated moving average (SARIMA) model with a linear model to forecast heat demand in a district heating system. They found that the linear model outperformed the SARIMA-model if weekly seasonality was included. Catalina, Iordache, and Caracaleanu (2013) used a multiple linear regression model on a simulated data set to predict future heat demand in Romania and found that this model performed well. Talebi, Haghighat, and Mirzaei (2017) used both an auto-regressive multiple linear regression and an auto-regressive multiple non-linear regression, and found that both models showed good agreement with results from more comprehensive modelling. Spoladore, Borelli, Devia, Mora, and Schenone (2016) used a linear regression with average daily temperature as only regressor to model natural gas demand at the town level and found reliable predictions. In addition, Tso and Yau (2007) compared decision trees, neural networks and regression analysis for predicting electricity consumption, and found results "indicating that the three techniques are generally comparable in predicting energy consumption". From the foregoing, it can be concluded that in numerous cases linear models are sufficient for the modelling and prediction of space heating demand.

### Machine learning methods

On the other hand, Zhao and Magoulès (2012) found that artifical neural networks (ANNs), which are machine learning methods, generally have a higher prediction accuracy than statistical models. Furthermore, Wei et al. (2018) found that "inaccuracy in short-term prediction and possible unforeseen correlations among the selected predictors greatly undermine the effectiveness of the regression models". In the same research, they present four other data-driven methods for prediction of building energy consumption. These methods are the artificial neural network (ANN), the support vector machine (SVM), the decision tree and the genetic algorithm. It concludes that among the data-driven approaches for prediction, ANNs perform best in a large number of applications ranging from load forecasting to retrofit potential estimation (Wei et al., 2018). In contrast, SVMs are relatively time-consuming for

large problems (Zhao & Magoulès, 2010), while decision trees are unable to deal well with time-series data, and besides are more suitable for predicting categorical variables (Tso & Yau, 2007). Amasyali and El-Gohary (2018) found that the ANN was the most used prediction method in their review of data-driven building energy consumption prediction studies, being the (main) research method in 47% of the studies they investigated. The ANN thus seems an appropriate method to improve the forecasting quality compared to a linear model.

Various types of artificial neural networks exist, and the ANN can be considered the standard neural network. The recurrent neural network (RNN) is designed specifically for time series data, as it 'saves' information from the former time step. Subsequently, it uses this information as one of the inputs in the next time step. Koschwitz et al. (2020) used a non-linear autoregressive exogenous (NARX) RNN to predict urban heating loads, and found that certain retrofitting orders affect the urban heating load. Talebi et al. (2017) used an ANN to develop a multiple non-linear regression model that included autoregressive components (this was a NARX model as well). In the latter research, the historical predictor values are used as predictors, while in the NARX-RNN of Koschwitz et al. (2020) the output of the RNN at $t-1$ is used as input for the RNN at $t$. A specific form of a recurrent neural network that has received much scientific attention lately, is the long short-term memory (LSTM) cell (Sherstinsky, 2020), developed by Hochreiter and Schmidhuber (1997). "The impact of the LSTM network has been notable in language modeling, speech-to-text transcription, machine translation, and other applications" (Sherstinsky, 2020). In general, LSTMs have been responsible for many of the best results from RNNs. Kong et al. (2017) used an LSTM to forecast the electric load of a single energy user, and found that "the proposed LSTM approach outperforms the other listed rival algorithms in the task of short-term load forecasting for individual residential households".

**Other methods**

From literature regarding space heating demand modeling, neural networks and linear models seem to be useful methods. These methods, however, are two extremes in the sense that a neural network is an extremely flexible and complex model, while linear models are relatively simple and restrictive. In order to 'bridge' this gap and find a method in between these extremes, the single index model and the functional coefficient model are looked into in more detail. These models combine aspects of a linear model with aspects of non-parametric models.

The idea behind the single-index model (SIM) is to extend the linear model by considering $x_i'\beta$ as the input for a function. As a consequence, the SIM takes the following form: $y_i = u(x_i'\beta) + \epsilon_i$, where $\beta$ has length $k$, which is the number of predictors. $u$ is a so-called link

function, and is thus a function of a linear combination of the predictors ($x_i'\beta$). Hence, the model assumes that it is possible to collapse the covariate $X$ to a single index (Wang, Xue, Zhu, Chong, et al., 2010). This seems like a strong restriction, but the model is still an extension of the linear model as it adds the function $u$. Cameron and Trivedi (2005) identified three possibilities to use the single-index model for panel data:

1. Additive individual specific effects model: $a_i + g(x_{it}, \beta) + \epsilon_{it}$
2. Multiplicative individual specific effects model: $a_i * g(x_{it}, \beta) + \epsilon_{it}$
3. Single index individual specific effects model: $g(a_i + x_{it}'\beta) + \epsilon_{it}$

While the SIM transforms $x_i'\beta$ with one function $u(\cdot)$, the functional coefficient model (FCM) transforms the linear model by making $\beta$ non-constant. The value of $\beta_k$ (the parameter for predictor $k$) depends on one or more other predictors (the varying variables). Changlin and Ning (2001) present the FCM as follows: $y = \beta_0(h(\mathbf{X})) + \beta_1(h(\mathbf{X}))x_1 + ... + \beta_k(h(\mathbf{X}))x_k$. Here $k$ is again the number of predictors and $h(\mathbf{X})$ is some function of the covariate $\mathbf{X}$. The predictors can be exogenous variables or auto-regressive components as done by Cai, Fan, and Yao (2000). The functions $h(\mathbf{X})$ do not have a certain specified form. These are simply estimated for certain values of the varying variables and subsequently interpolated.

The SIM and FCM are thus small extensions of the standard linear model. Because of the extra flexibility the models offers they can improve (prediction) results compared with the linear models. On the other hand, they are much simpler than artificial neural networks, which results in a more time-efficient estimation and more insight in the structure of the relations in the data.

### 2.1.2 Predictors

There are two main categories of variables that affect energy demand for heating or cooling within a building: climatic variables and the buildings' (architectural) properties (Dolinar, Vidrih, Kajfež-Bogataj, & Medved, 2010). On the other hand, "occupant behavior is the greatest uncertainty in building energy consumption prediction" (O'Brien & Gunay, 2015). In practice, however, occupant behaviour is rarely included in space heating demand models, as it is very difficult to measure.

The most important climatic predictor is temperature (Fang & Lahdelma, 2016). Especially with a low temporal granularity, degree days (DD) are generally used instead of temperature. Degree days are computed using equation 1, where $T_{thr}$ is the threshold temperature, and $T_i$ is the temperature at day $i$.

$$DD = \sum_{i=1}^{T} \max(T_{thr} - T_i, 0) \tag{1}$$

After computing degree days for one day, all degree days are summed for the time period 1 to $T$. The threshold for degree days depends on the quality of insulation (Verbai, Lakatos, & Kalmár, 2014). Note that here, the idea for heating degree days is explained - for cooling degree days the counting starts when the temperature reaches above a certain threshold.

Jylhä et al. (2015) compared the degree days method with a dynamic building energy simulation tool, using hourly data for modelling heating and cooling demand. They found results that suggested that the heating degree day method would work adequately to predict heating demand, although for April and September the decrease in heating demand was underestimated. Corrections or weights for the degree days of certain months or seasons, as used by Spoladore et al. (2016), could improve such results. Cox, Drews, Rode, and Nielsen (2015) found that "even coarse annual estimates of air temperature change produce useful estimates of energy demand", indicating that using detailed weather input is not always necessary. Verbai et al. (2014) found that heat islands can have important effects on heat demand in cities, and that degree day values should be corrected accordingly.

Looking at the building characteristics, Dascalaki, Droutsa, Balaras, and Kontoyiannidis (2011) mention that building typology is an important variable to predict the energy performance. Different typologies are e.g. detached houses, semi-detached houses and terraced houses. Regarding building properties, Koschwitz et al. (2020) argue that building refurbishments, where insulation is improved, will highly affect future energy demand of buildings. Frank (2005) also found that the thermal insulation level has an important effect on heating energy demand. In the same research, it was concluded that the building year affects the thermal insulation level, because in general newer buildings have better insulation levels than older ones. Lopes, Hokoi, Miura, and Shuhei (2005) included the floor surface and building year as well. They found a negative relation between building year and the heat transfer coefficient. The predictor floor surface is used as an indicator of the air volume that needs to be heated.

Talebi et al. (2017) and Brown et al. (2012) used demand at $t-1$ as a predictor, i.e., these researchers used an auto-regressive model. This could help to model the occupant behaviour, although an individual intercept $\alpha_i$ might be a more elegant solution. For some readers, it might be surprising that price is a predictor that rarely is used. However, heat can be considered a utility of which the price barely influences the demand. In their meta-analysis, Labandeira, Labeaga, and López-Otero (2017) classified energy products as price inelastic.

In addition, they found that "the least price-sensitive energy good is heating oil both in the short and long term" (Labandeira et al., 2017). As a consequence, the heat price is not taken into account in this research.

### 2.2 Temperature and degree days

To model future temperatures, it is common to use predictions of some climate modelling institute (such as the International Panel on Climate Change or a national weather institute). As such institutes often present predictions for specific years, interpolation between their predictions is used to obtain predictions for every year or month. This is for example done by Andrić et al. (2016) and Koschwitz et al. (2020). Sometimes temperature reference year is used, but this method is not suited for modelling extremes in weather (Frank, 2005). Spinoni et al. (2018) predicted the change in heating and cooling degree days for Europe until 2100. They found an annual decrease in heating degree days for the Netherlands of 5 to 7 days, which resulted in a predicted decrease in heating degree days between 200 and 500 in the period 2040 to 2070.

### 2.3 Insulation and energy labels

The European Union obliged member states to assign an Energy Performance Certificate to houses when built, sold or rented (European Parliament and Council, 2002). These Energy Performance Certificates are now known as energy labels, ranging from A (very energy efficient) to G (very energy inefficient). A research conducted in the Netherlands has shown that a better energy label indeed reduces energy demand, but that the reduction in energy demand was smaller in practice than in theory (Majcen, Itard, & Visscher, 2013). The energy label is used as a proxy for insulation quality in this research. However, it thus has to be seen if the energy label is indeed a good proxy for the insulation level. In addition, it needs to be investigated to what extent insulation improvements to residential buildings will be made in the future. To the author's knowledge, such an investigation has not been conducted yet.

### 2.4 Forecasting space heating demand

Few scientists found clear results regarding the average future space heating demand of households. Andrić et al. (2016) used a resistance-capacitance model (based on principles of thermodynamics) including hot tap water demand and space heating demand. They predicted a heat demand decrease between 22.3% and 52.4% in 2050 for a district in Lisbon. As the hot tap water demand is expected not to decrease, the predicted decrease in space heating demand is actually even larger than between 22.3% and 52.4%. They also found results that implied that the building renovation could have higher impact on future heat demand than changes in weather parameters. Dutch researchers predicted a yearly decrease of 1.5% for the

Netherlands (Segers, Van den Oever, Niessink, & Menkveld, 2019), mentioning climate change and energy savings as causes. This would result in a decrease of 36.5% in 2050 compared to 2020.

In order to yield lower forecast errors, Bates and Granger (1969) proposed to combine different sets of forecasts. Timmermann (2006) stated that "forecast combinations have frequently been found in empirical studies to produce better forecasts on average than methods based on the ex ante best individual forecasting model". Estimating the 'best' combination of different forecasts has since been an important challenge for scientists. However, "empirical evidence and extensive simulations show that the estimated optimal forecast combination typically does not perform well, and that the arithmetic mean often performs better" (Claeskens, Magnus, Vasnev, & Wang, 2016). Smith and Wallis (2009) presented a formal explanation why simple combined point forecasts "are repeatedly found to outperform sophisticated weighted combinations in empirical applications". This explanation, however, is not discussed in more detail here.

To conclude, this research addresses a research gap by investigating the space heating demand of households in the Netherlands using (multiple) econometric methods and combined forecasting. In addition, as earlier stated, it can contribute to the limited insights in predicting space heating demand (i) with high spatial granularity, (ii) for residential areas and (iii) with a long-term horizon.

# 3 Data

In this section the available data is described. First, the main conclusions from the data exploration are presented (section 3.1). After this short section, this chapter contains a part dealing with space heating demand data (section 3.2). Hereafter, in section 3.3, the steps in the data preparation are discussed and the first results hereof are presented. Then, temperature data (section 3.4) and energy label data (section 3.5) are described. The temperature and energy label data are necessary to make forecasts of the future space heating demand, assuming that these factors influence the space heating demand. All heat demand data is provided by Eneco and based on the yearly bills of their clients. The forecasts for temperature and energy labels are combined with the historical building data to scenarios in section 3.6.

## 3.1 Data exploration

The data exploration consists of three main parts. For a more detailed description of the data exploration, readers are referred to appendix A.2.

First, it is found that the hot tap water demand is unrelated with degree days. Degree days are computed for 11 different years. The relation between degree days and hot tap water demand is investigated using linear regression, with degree days as only regressor to predict hot tap water demand. The degree day coefficient is highly insignificant (p= 0.736), and it can thus be concluded that it has no effect on hot tap water demand.

Secondly, the relation between temperature and total heat production is assessed. The heat production minus transportation losses is equal to the heat demand. It is found that there exists a base load demand, which seems fairly constant. Heat demand increases for lower temperatures, and starts to grow below certain a threshold temperature. Naturally, there is not an exact threshold (this differs per household, depends on the amount of sunlight, etc.).

Thirdly, this threshold is investigated. Degree days based on average and maximum daily temperature function well to predict the total heat production. Averaged $R^2$-values over 13 regressions (for different locations) are all above 0.8 for the seven maximum and average temperatures investigated. Using a minimum temperature as threshold functions clearly worse. Based on these $R^2$-values, using the maximum temperature with a threshold of 19 degrees Celsius would make most sense. However, it is common to work with average daily temperatures. More specifically, it is easier to find average temperature predictions than maximum temperature predictions. As 16 degrees Celsius is the best predictor using average temperature, the threshold used in this research is an average temperature of 16 degrees Celsius.

## 3.2 Space heating demand

The space heating data set contains 23 variables for 108,036 households. All bulk consumers are excluded. The 23 variables can roughly be split in two parts. First, there are 11 building-related variables and secondly there is the household space heating consumption per year from 2007 until 2018. The data set is hence a panel data set: for the households (the observations), there is data available for different years. In Appendix A.1, a list with an explanation per variable is presented.

### 3.2.1 Missing data

Figure 1 shows the data availability per variable. It makes clear that the space heating demand availability increases from 2007 to 2018. The building characteristics are almost fully available, except for the energy label data. About half of the energy label data is missing.

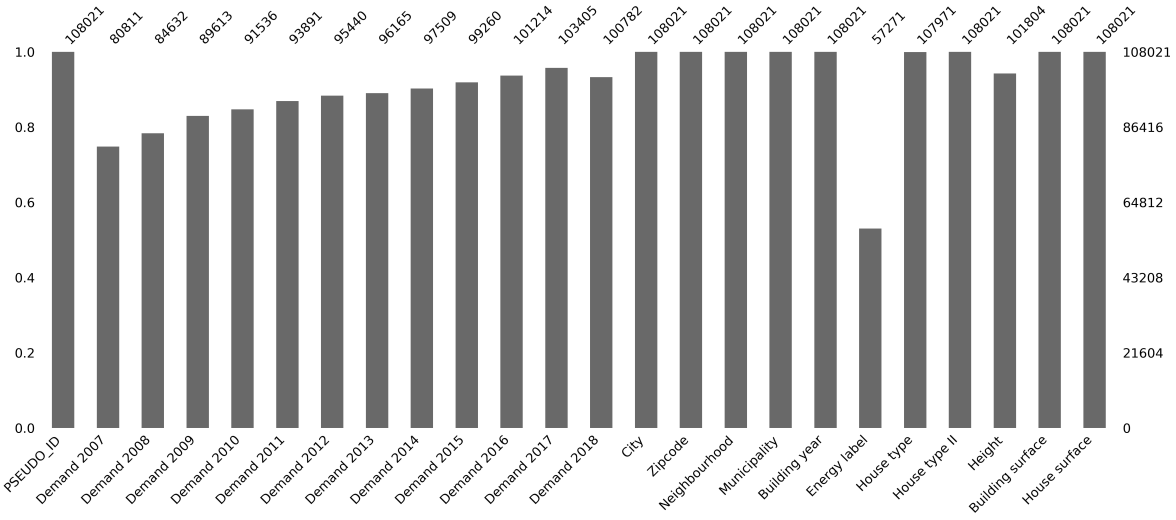


Figure 1: Absolute number of data points available per variable.

Figure 27 in appendix A shows correlations between the availability of variables. There are positive correlations between space heating demands in sequential years. This could be expected: if a household was not an Eneco client in 2008 it is likely that it still was not in 2009. In addition, there are positive correlations between the availability of height class data and the availability of demand data. This means that if a household's height class is known, the chances are higher it has been an Eneco client from earlier on.

### 3.2.2 Descriptive statistics

The data set contains a considerable amount of outliers, especially regarding the space heating demand. It is rare for households to exceed a yearly space heating demand of 200 gigajoule

(GJ), and the mean and median in the data set are between 32 and 66, and 15 and 33 respectively (before data preparation, see table 7 in appendix A.3). However, the available data contains values of above 10.000. The means are affected by these outliers and might thus actually be lower. Because of the outliers, boxplots are hardly readable (see figure 25 in appendix A.3). A bar plot, as given in figure 2, hence gives a clearer picture. It shows that the space heating demand data is skewed, with the majority of the values between 0 and 40, but containing values above 1000 as well.

Other bar plots in figures 3, 4 and 5 give clarity about the characteristics of the buildings in the data set. These are all Eneco clients, and this is thus not per definition representative for the Netherlands. Most of the buildings are built between 1981 and 1990, or between 2001 and 2010 (see figure 4). There are barely any buildings built before 1950. Most of the houses are either in cities (Utrecht, Rotterdam and The Hague) or suburbs (Nieuwegein, Capelle aan den IJssel and Houten) (see figure 5). The great majority of the houses is either a terraced house or a house in a flat (see figure 3).

When looking into the demand for specific subsets of the data, it becomes clear that there are large differences even for houses with (very) similar characteristics. This is even the case when some of the main predictors have exactly the same value. Figure 6 shows that even in the warmest year (2014, with only 1778 degree days), for recently built flats (in 2006) with the energy label that indicates the highest efficiency (label A) a large spread in space heating demand exists. This set is specifically selected to have a small spread, because all characteristics are expected to result in a low demand. However, the 163 houses with these characteristics have a demand ranging from 0.3 to 70.6 GJ, and a variance of 156.7. The range for this small subset, thus specifically designed to have little dispersion, spans almost a quarter of the range of demand in the complete data set, while containing only 163 out of more than 900,000 data points.

Figure 7 shows the results for a subset that is designed such that high demand values are expected. It contains data from the coldest year (2010), for relatively old houses (building year 1984). The energy label (B) and the house type (terraced) are chosen in order to get a substantial subset of 108 households. The values are indeed higher, ranging from 11.6 to 133.4 GJ. The variance is a lot higher as well, 401.2. Although in this case the variance is strongly affected by two values higher than 120, the range and variance would still be large when these two values would be excluded. The maximum value without these two extremes is 96.4, and the variance without them is 256.1.

The statistics for demand for these two subsets are not in solitary, but naturally not all

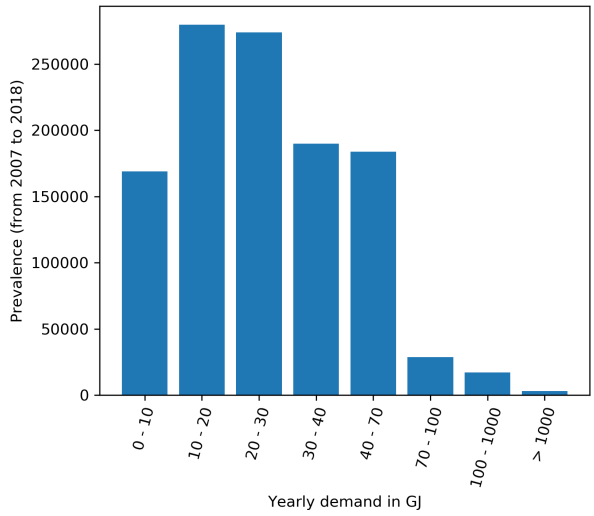


Figure 2: Prevalence of space heating demand values for 2007 - 2018.

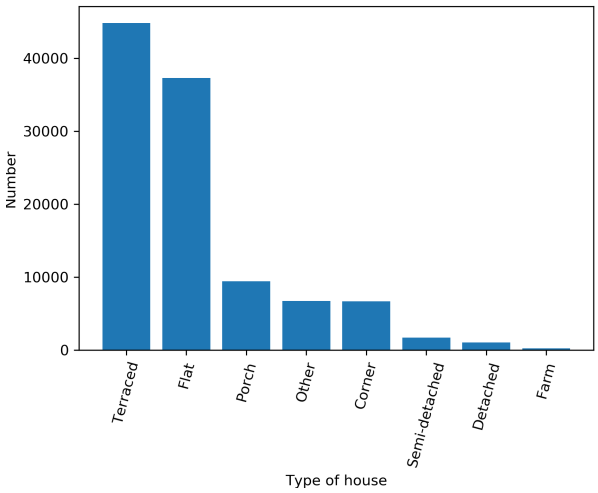


Figure 3: Types of houses in the data set.

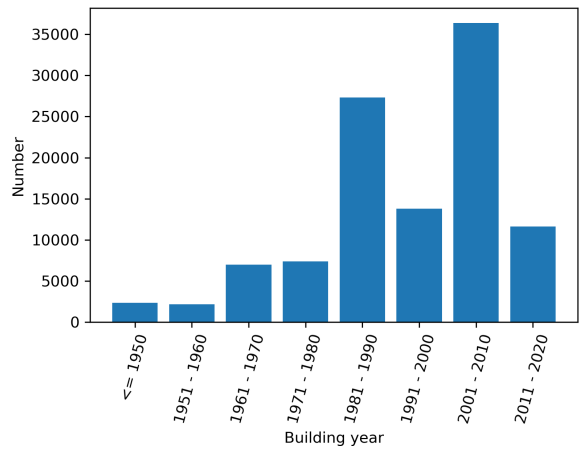


Figure 4: Building years of houses in the data set.

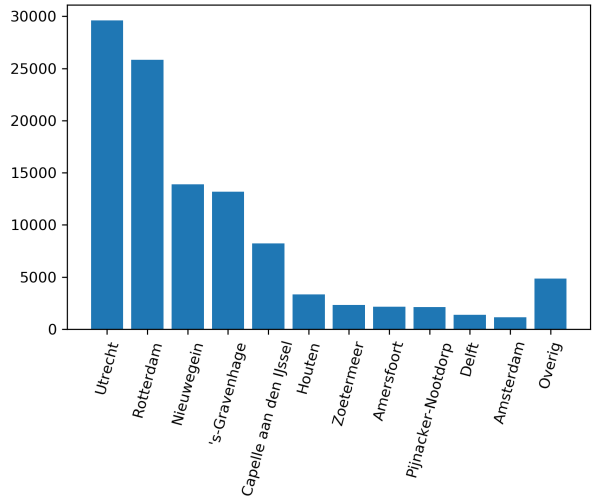


Figure 5: Houses per municipality.

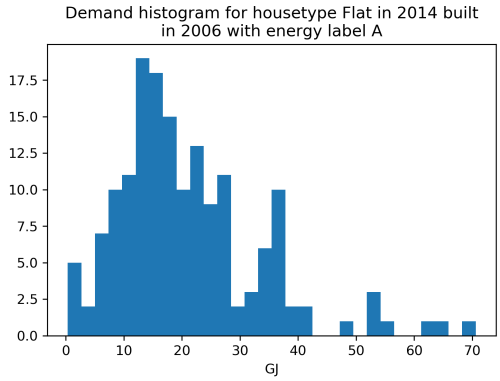


Figure 6: Histogram for demand values for flats, built in 2006, with energy label A for the year 2014.

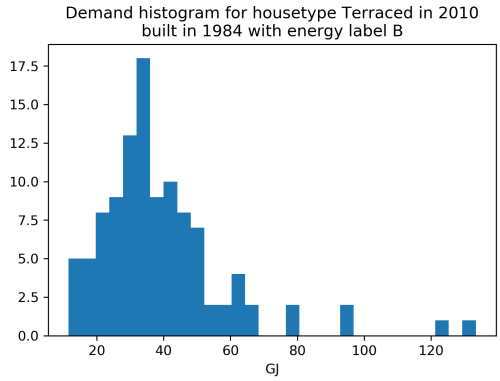


Figure 7: Histogram for demand values for terraced houses, built in 1984, with energy label B for the year 2010.

examples are shown here. These results, however, indicate that there is much noise within the data set, or that some important factor is not included in the predictors. Such a factor could be occupant behaviour. Note that floor surface is also excluded to keep sets of substantial size. One could however expect that for both the house types flat and terraced house, extremely high floor surfaces are unlikely. As a consequence of what is found in the foregoing two paragraphs, it is expected that it will be difficult to obtain good results for predicting the demand of individual households.

## 3.3 Data preparation

A cleaner data set can improve the results, and hence "data pretreatment is another focus for many researches" (Wei et al., 2018). The data preparation consists of dealing with 'incorrect' outliers, data imputation and dealing with correlation outliers. In addition, some practical preparations are needed to get the data in the correct form for different methods. These steps, however, are not explained into further detail. Lastly, the splitting of the data set in training, validation and test set is explained.

### 3.3.1 Incorrect outliers

Two types of outliers are considered. First of all, some data points contain 'incorrect' values. This can of course never be known for sure, but there might be serious reasons to suspect this. In this research, a yearly space heating demand higher than 300 GJ is considered an incorrect outlier, because it is considered an unrealistically high value for a household (Fekkes, 2020). These demand values are deleted, and subsequently imputed. The second type of outlier is a so-called correlation outlier. These are dealt with in the third step of the data preparation, using the Minimum Covariance Determinant.

### 3.3.2 Data imputation

To determine what to do with missing data, the reasons why data is missing have to be examined. For this research, the missing data is considered missing completely at random (MCAR). For the explanation about missing data mechanisms and the argumentation why the missing data is MCAR, readers are referred to Appendix A.4. Although the missing data is MCAR, data imputation is applied in this research, to keep as much valuable information as possible.

Data imputation is the operation of estimating values for missing data. To give a basic example, one of the simplest options is to impute the mean value for a certain variable. There are many more ways of imputing. It is possible to impute either a single value, or multiple values (this is a different decision parameter than the way of imputing). In single imputation, imputing one value is considered sufficient. This method takes into account the variability in the data available, but not the variability that could be caused by the missing data. Multiple imputation imputes multiple values (and results thus in multiple data sets). As a result, this method can include variability from the missing data.

In this research, single imputation is used to impute the missing data. For the space heating demand, relatively small percentages of the data are missing and the missing data is considered MCAR. Hence, it is unlikely that a true part of the variance misses. This makes it likely that single imputation suffices as imputation method, because the extra variability from the missing data is limited. For the energy labels, all the possible values are known - A to G - and they are all in the data set. A larger part of the data is missing, but because the range of possible values is known, again, single imputation seems to be sufficient.

The data imputation is done in two different ways. The first method deletes data points that contain missing demand data (for any year) and uses model-based imputation to impute other missing data. This results in data set 1. The missing data demand data could be deleted, as it is considered missing completely at random (again, see appendix A.4). The second method uses model-based imputation for all missing data and results in data set 2. Table 1 presents a comparison of how missing data was dealt with per variable that contained missing data for the two different data sets. During the research, the data sets are compared on their prediction quality. The data set that performs best in prediction quality is used for forecasting.

The model-based imputation was executed using the Python *scikit* package (Pedregosa et al., 2011), using a linear regression for numerical variables. This imputation was executed iteratively, starting at the mean and using 10 iterations. The same Python package was used to impute categorical variables, where the mode was imputed.

### 3.3.3 Minimum Covariance Determinant

Having deleted outliers that were considered incorrect, and having imputed data, it is time to look at correlation outliers. The first method of outlier deletion could not detect (and thus not delete) correlation outliers. A correlation outlier does not have an extreme value on one variable, but is still an outlier because of the combination of non-extreme values (such as the red data points shown in figure 8). These outliers can be detected using the minimum covariance determinant (MCD) estimator (Rousseeuw, 1985). In short, the MCD tries to find the group of data points that has the smallest covariance matrix determinant. It is a way to minimise the covariance of the chosen sample. As the covariance is a measure of spread, this is a way to minimise the spread. As such, outliers are excluded from the set. Because the MCD looks at the statistical distance, it takes into account that some points are statistically closer to the mean than other points that might be closer in absolute terms.

The MCD is also able to detect outliers that contain correct values, but that have a large statistical distance from the rest of the data points. These data points are sometimes deleted as well because they affect the estimates.

Computing the MCD and the outliers was done using again the *scikit* package by Pedregosa et al. (2011). The MCD was determined considering all data of a household as one observation (so including the demand for all years available). As a consequence, a household was excluded or included as a whole to keep a balanced panel data set (i.e., with the same amount of observations per household).

However, deleting the data points that were considered outliers by the MCD, would harm

Table 1: Differences of the data sets in imputation method and size.

| | Data set 1 | Data set 2 |
|---|---|---|
| Demand | Deleted | Linear model |
| Building type | Mode | Mode |
| Energy label | Linear model | Linear model |
| Households | 75912 | 108021 |

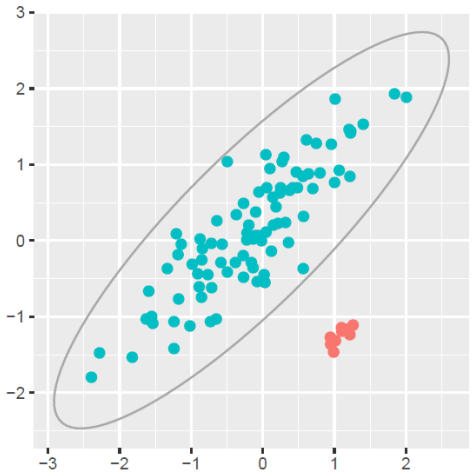


Figure 8: Correlation outliers in red, with the non-outliers in blue (Alfons, 2019).

Table 2: Energy label and demand statistics for the two data sets before and after applying the MCD. For energy labels, value 1 corresponds with energy label A (most sustainable), and value 7 corresponds with label G (least sustainable). For data set 2, there are some extremely high values imputed for demand. This is the result of the imputation of demand values for houses with large floor surfaces.

| | Energy label | | | | Demand | | | |
|---|---|---|---|---|---|---|---|---|
| | D1 | D2 | D1 MCD | D2 MCD | D1 | D2 | D1 MCD | D2 MCD |
| Mean | 2.56 | 2.40 | 2.47 | 2.30 | 28.88 | 28.28 | 24.56 | 24.03 |
| Median | 3.00 | 2.00 | 2.00 | 2.00 | 25.02 | 24.57 | 22.86 | 22.61 |
| Max | 7.00 | 7.00 | 5.00 | 5.00 | 299.75 | 1986.52 | 103.35 | 102.17 |
| Min | 1.00 | 1.00 | 1.00 | 1.00 | 0.01 | 0.01 | 0.01 | 0.01 |
| SD | 0.97 | 1.05 | 0.79 | 0.85 | 21.08 | 22.12 | 13.86 | 13.17 |

the variation in the predictors. The energy labels F and G (corresponding to numbers 6 and 7) are excluded by the MCD (see table 2). Furthermore, the range in demand decreased from roughly 300 to 100, excluding almost all demand values higher than 100 GJ (see table and 2). It is possible to set the proportion of data points to include in the first iteration of computing the MCD. However, even for very high values of this so-called support fraction (e.g. keep 99% of the data in the first iteration), an important part of the data was considered an outlier. As a consequence, the MCD outliers are not excluded from the data to keep the variety present in the data set. The data sets obtained after imputation are thus used, including the data points considered correlation outliers by the MCD. The likely cause of the malfunctioning of the MCD is the fact that the data is non-symmetric (see for example figures 25 and 26 in appendix A.3), while the MCD is not designed for non-symmetric data.

### 3.3.4 Splitting the data

The data set is split in three sets: the training set, the validation set and the test set. The training set is used to optimise the different models for the different methods. Subsequently, the validation set is used to compare different models within a method. Finally, the best models per method are compared using the training and validation set together for fitting, and the test set for out-of-sample prediction. The test set is not included in fitting the models yet, so that the method comparison can be completely fair. The idea behind this is that if the test set is included in choosing the best model, it can already have influenced the settings for the model.

It is important to note that the splits are made over households. The validation and test set thus contain households that are not included in the training set. On the other hand, all years are included in every set. The split in this direction is chosen because there are many households included (see table 1), while there are only 12 years. Splitting the data over time would thus strongly decrease the spread in years, and would also strongly decrease the variety of degree days in the training, validation and test set. As a result, the structural differences between the different sets are smaller for a split over households.

The used split sizes differ for the two data sets, as they differ in size. Split sizes of about 60/20/20 or 70/15/15 percent (in the order training, validation, test set) are recommended by Ng (2020) for data sets that are not extremely large (extremely large meaning having multiple millions of data points). In addition, the number of hyperparameters that has to be estimated is not extremely large (e.g. four for the RNN), and as a result a relatively small training set should suffice. Hence, data set 1 is split in the sizes 60/20/20 percent. For data set 1, the training set contains 45548 households and the validation and test set contain 15182 households. Naturally, the same validation and test set for data set 2 are used. As a consequence, the training set contains 71.8% of the data points and the validation and test set both contain 14.1% of the data points for data set 2. The training set of data set 2 contains 77657 households.

However, in the method comparison, different splits are tested to check the validity of the results (and of the split sizes). As the data is of substantial size (i.e., having more than hundreds or thousands of data points), larger training sets and smaller test sets are used. Different split sizes are used when comparing the methods, because the differences between methods are expected to be bigger than the differences between different model of one method. To give an example, the difference between the best performing RE model and the best performing RNN model is likely to be larger than the difference between RE with or without one

predictor, or RNN with one node per layer more or less. As such, testing different data splits in this stage of the study is more relevant than testing different data splits in an earlier stage.

The methods are compared on prediction quality for the test set. In this part of the study, 80 % (the training and validation set combined) of the data are thus used as the training sample (for data set 1). To validate the split sizes, training sets of sizes 85%, 90% and 95% are tested as well. The splits are made randomly again for the new split sizes, i.e. the training set of size 85% does not contain the all data points of the 80% training set and then 5 percentage point extra data points. This increases the chances of a different results and is as such a more critical test for the validity of the split sizes.

### 3.4 Temperature data

Degree days is the only predictor used that is related to weather. Fang and Lahdelma (2016) found that wind effects are small compared to temperature effects. In addition, the Dutch Royal Meteorological Institute (KNMI) climate scenarios demonstrate that barely any change in wind speed in the future is expected (Klein Tank, Beersma, Bessembinder, Van den Hurk, & Lenderink, 2014). Solar radiation is more often excluded than included in space heating demand models. Furthermore, the KNMI climate scenarios again expect very little change in this variable compared to the reference period: for all scenarios the relative change is between -0.8% and +1.6% (Klein Tank et al., 2014).

The historical temperature data (for the period 2007 to 2018) is attained from the KNMI (2020a). The temperature data from their main weather station in De Bilt, which is also (one of) the most central weather stations, is used. From the daily average temperature, the degree days per year are computed, using equation 1. As yet stated, the threshold $T_{thr}$ is 16 degrees Celsius.

The KNMI also made forecasts for the average temperature in the Netherlands. For every year that is a multiple of five, temperatures are forecast. For every one of these forecast years, thirty years exemplary years are used, all of which are considered equally likely. To give an example, for the year 2040, daily mean temperatures are predicted for thirty different years. Historical temperature data from 1981 until 2010 function as a reference period. The idea of predicting thirty years of mean daily temperatures is that this includes the variability in temperature that can occur. So if you would like to know the prediction of January 1st in 2040, you can look at the predictions for January 1st for thirty years. Based on this set the expected temperature and its variability can be determined.

These predictions are publicly made available (KNMI, 2020b). The KNMI has developed four climate scenarios containing future temperatures, based on its own research and research from the Intergovernmental Panel on Climate Change (IPCC). The IPCC developed two worldwide climate scenarios, one with a moderate temperature increase (scenarios with 'G') and one with a strong temperature increase (noted with 'W'). The second aspect included in the climate scenarios is the change in air stream pattern in the Netherlands. This pattern could have a low impact (noted with 'L') or a high impact (noted with 'H') on the Dutch climate. All combinations for the two aspects are possible, which results in the four climate scenarios GL, GH, WL and WH. Figure 9 shows the expected number of degree days per year for the two most extreme scenarios and the boundaries of the 90% confidence interval for these scenarios. The figure contains historical data as well. These data place the the KNMI predictions in some perspective. It is, however, important to note that very few historical data points are shown, and it is thus incorrect to draw (strong) conclusions from these historical data points. Because of the variability included in the prediction, as described in the former paragraph, it is possible to compute standard deviations for these expected number of degree days too. The estimates of average yearly decrease in degree days from the KNMI range from 5.3 in the GL scenarios to 11.5 in the WH scenario. Compared to the predictions of Spinoni et al. (2018), who predicted an annual decrease between 5 and 7 degree days for the Netherlands, the range of the KNMI is wider and the estimated decrease is higher.

The differences between the scenarios in 2020 can be explained by the fact that the scenarios stem from 2014. As a result, the predictions already differed in 2020. One might also note that the standard deviations do not increase for the scenarios (but decrease slightly in absolute sense). An increase in the standard deviation could be expected arguing that things further in the future are more insecure. However, the increasing insecurity is included in by the increasing differences between the scenarios (as is clearly visible in figure 9). In addition, the standard deviations do increase relatively to the mean.

### 3.5 Energy label data

As yet mentioned, energy labels are used as a proxy for insulation level. Energy labels are standardised within the European Union (European Parliament and Council, 2010). The National Service for Entrepreneurs (RvO in Dutch) keeps track of all energy labels in the Netherlands since 2009. This data is freely retrievable from their website (Rijksdienst voor Ondernemend Nederland, 2019) and is used as a basis for making scenarios. The left part of figure 10 shows the percentage of houses in the Netherlands with a certain energy label from 2009 to 2018. The right part consists of estimations. 46% of the Dutch houses had an energy

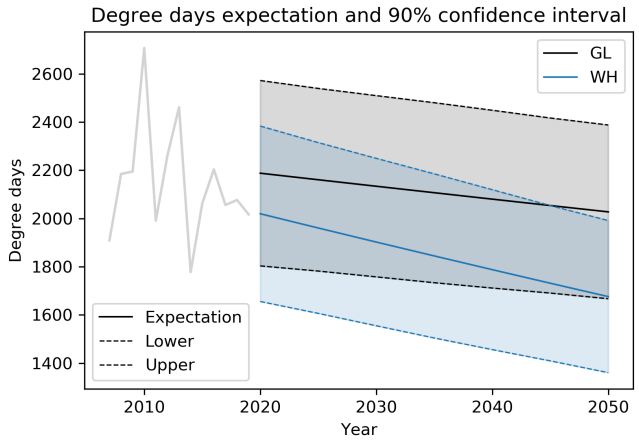


Figure 9: The number of degree days in two most extreme climate scenarios, GL and WH, and their 90% confidence intervals. The 90% confidence intervals are shaded in the corresponding color.

label in 2018 (this is the most recent number available) (Rijksdienst voor Ondernemend Nederland, 2020).

It is unrealistic to assume that energy labels up to 2050 (30 years ahead) can be predicted using data that describes only ten years. This is especially difficult as it is subject to local, national and international governmental policies. Hence, four scenarios are developed to see the effects of a range of future possibilities. The scenarios are developed in line with the Eneco scenarios called Existing policy, Accelerated policy, Circles and Tides. A short description of these scenarios is given in appendix A.5. This includes a detailed description of the energy label scenarios.

A general assumption is that the energy labels of the housing stock will improve. This will be the case even with few renovations because newly built houses have higher insulation quality requirements than older houses (Rijksoverheid, 2020c). Figure 10 shows the scenario for existing policy, including the historical data. Only the circles scenario reaches a full housing stock with A-labels in 2050. The idea is that in all scenarios this situation will be achieved in the future, albeit with different paces. Figure 11 compares the scenarios on the percentage of A-labels, showing these different paces.

### 3.6 Scenarios

The scenarios are developed using the historical data for the building characteristics, and both the climate and energy label scenarios. The basis is formed by the available information

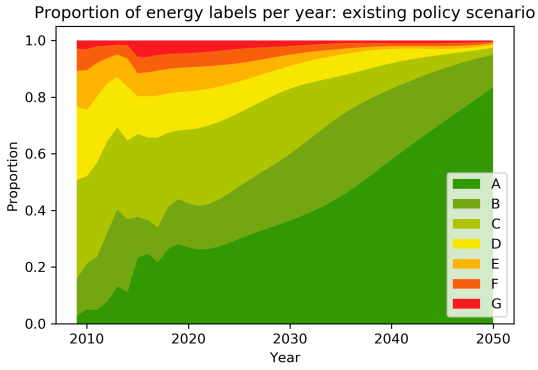


Figure 10: The proportion of energy labels in the existing policy scenario.

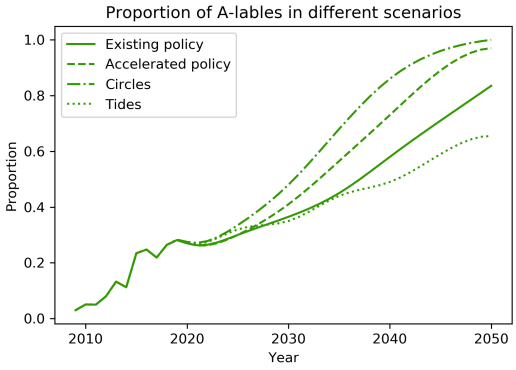


Figure 11: The proportion of A-labels in the different energy label scenarios.

about the housing stock. Thus, the building year, floor surface, and housing type of the housing stock from the historical data set are used. For every year, energy labels are randomly assigned to houses, using the proportion of energy labels corresponding to the year and energy label scenario. Next, the expected degree days per year are added. This results in sixteen main scenarios (as there are four main climate scenarios and four energy label scenarios). These scenarios, however, do not provide the variation within climate scenarios.

To allow for variation within climate scenarios, the data made available by the KNMI is used to the full. As earlier explained, for every future year they provided 30 possible years with daily average temperatures. All these thirty years are considered an equally likely possibility. Thus for every year, there are 120 possibilities (30 years for four scenario). For all these 120 possible years the degree days are computed. These 120 options are combined with the four energy label scenarios and this results in 480 possible data sets for one year. These 480 sets are used to produce the boxplots shown in figure 17 in section 5.3.

To present an idea of the variation within one climate scenario, confidence intervals are produced. To compute the boundaries of these confidence intervals, the 30 possible future years (that corresponds to one year) are used. A confidence interval of 90% is used and results for the boundaries of these confidence intervals are presented in section 5.3. These scenarios are called the 90% scenarios. In total, there are 32 of them, presenting an upper and lower boundary for all 16 main scenarios.

# 4 Methodology

In the first section of this chapter, some assumptions of this research are stated explicitly. In the second part, the three methods for modelling the space heating demand are clarified. In part three, the evaluation OF different methods is explained. In the fourth section, the combination of forecasts is discussed. In the last part, hypotheses for the research questions are presented.

## 4.1 General assumptions

In this research the number of clients of a district heating network is not included. As yet discussed, it is about finding the (change in) space heating demand per household. Hence, factors that can affect the number of households connected to a district heating network are not included. Furthermore, changes in building characteristics are not included, apart from changes in energy label (see section 3.5). Hence, the fact that Eneco's portfolio in terms of household or building characteristics might be different in the future, is not taken into account.

This research is specifically about space heating demand, and not about cooling demand. Cooling demand depends on the climate and insulation as well, but cooling demand is in the Netherlands not supplied via district heating networks.

A third assumption is that heat demand does not depend on the heat price. This is because people need heat to stay warm, and this is considered such a necessary good that it is price-inelastic. In addition, heat prices in the Netherlands are regulated to keep them affordable.

## 4.2 Descriptive model for space heating

From the literature, it became clear that linear models generally perform considerably well for forecasting space heating demand. In addition, these models are useful for interpretability. As the data set has a panel structure, the statistical method in this research is a linear panel data model (section 4.2.1). The assumption of linearity is of course a strong one. However, such models are usually a good starting point for understanding the data. Moreover, it is possible for some of the linear panel models to add an intercept that differs per household. This could be very useful to model the occupant behaviour.

To increase the flexibility, the functional coefficient model (FCM) is used to model space heating demand as well, dropping the assumption of linearity. This model is preferred over

the single-index model because it is expected to suit the data better. The expectation is that degree days could function very well as the *varying variable*. Imagine it is 30 degree Celsius outside, and there are thus zero degree days. In this case, it is unlikely that any other predictor causes space heating demand, no matter how badly your house is insulated, or how large the floor surface is. On the other hand, on very cold days, it is likely that, for example, insulation quality results in a higher reduction of space heating demand than on days with temperatures just below the degree day threshold. This indicates that the parameter for all predictors might depend on the value of degree days in a certain year. The functional coefficient model relates the predictors and the demand exactly in such a way. The varying variable is thus called as such, because it varies the parameter of other predictors. In addition to the expectation that the FCM fits the data, the single-index model is only a very small extension to a linear model. As a consequence, the results of the single-index model and the linear model could be very similar. The functional coefficient model is discussed in section 4.2.2.

The third and last method used in this research is a recurrent neural network (RNN). The RNN is developed to investigate the performance of a very flexible method on these data and to investigate the inclusion of a dynamic component. The method is presented in more detail in section 4.2.3. As the long short-term memory (LSTM) cell is the RNN variant responsible for the many of its good results, both the regular RNN and the LSTM are included in this research. The inclusion of a dynamic effect could contribute to modelling changing inhabitants of a building. When inhabitants change, this cannot be modelled by a household-specific effect as in some of the linear panel data models, because such an effect is constant over time. However, a dynamic effect can do this, although only partially. A possible result of the dynamic effect is that the space heating demand at $t$ is related more strongly to the space heating demand at $t-1$ than at $t-n$ for $n > 1$. This captures that a building is more likely to have the same inhabitant at $t-1$ than at $t-n$. The idea is that if it is uncommon to move, and people thus inhabit a house for a long time, there is a strong dynamic effect. In such a case, the space heating demand at $t$ is likely to be similar to the years before, as the building had the same inhabitants. If the inhabitants of a house change often, the previous years are not such a good predictor, and the dynamic effect is likely to be weak.

The results of using the aforementioned methods, is that a balanced set of methods is applied to the data. This should result in a reliable estimate when combining forecasts of the methods.

### 4.2.1 Linear panel data models

The most basic linear panel data model is the so-called pooled ordinary least squares (OLS) model. The formula for this model is given in equation 2, where $y_{it}$ is the space heating demand of household $i$ at time $t$. It is fitted on an intercept $\alpha$ and the set of independent variables $x_{it}$ (including both time-related variables and building-related variables). $\beta$ is a vector of $k$ parameters, where $k$ is the number of predicting variables in $x_{it}$. $\epsilon_{it}$ is the error term for observation $y_{it}$. In all linear panel data models, categorical variables are modelled using dummies. The energy label is modelled as an interval variable (assigning 1 to label A, 2 to label B etc.).

$$y_{it} = \alpha + x'_{it}\beta + \epsilon_{it} \tag{2}$$

A relevant extension is including a specific effect for every household. The idea of using the specific effect is to model the occupant behaviour, which from the literature appeared to be a difficult aspect to model. The model equation is given in equation 3. In this model, a household-specific intercept $\alpha_i$ replaces the general intercept $\alpha$ from the pooled OLS model.

$$y_{it} = \alpha_i + x'_{it}\beta + \epsilon_{it} \tag{3}$$

The intercept $\alpha_i$ is thus specific for all households. This household-specific effect can be fixed or random. A fixed household-specific effect $\alpha_i$ can be correlated with one or more of the regressors $x_{it}$, while a random household-specific effect $\alpha_i$ is assumed to be uncorrelated with all regressors $x_{it}$. In our case, the occupant behaviour could for example be correlated with the floor surface of the house. A family with three children is likely to have a higher space heating demand than a couple, and it probably lives in a house with a larger floor area too. Hence, a correlation between floor surface (a regressor) and occupant behaviour (modelled in the household-specific intercept) could be expected. On the other hand, for many other regressors a correlation with occupant behaviour is unlikely, such as the building year of the house. In theory, it is thus unclear which model would perform better. The fixed-effect model, however, cannot be estimated when regressors are included that vary only over individuals (households in this case) and not over time (Naghi, 2018). This is caused by the assumed relation between regressors and the household-specific intercept in the fixed effect model. The data set contains many of such regressors (like building year, floor surface and the energy label), which can thus not be included if the effect is assumed fixed. As a result, the random effects model is expected to perform better than the fixed effect model. Still, both models will be tested and compared, as it can provide extra insight in the data. The fixed effect (FE) model is estimated using the within-estimator. The random effects model (RE) is estimated using Generalized Least Squares.

The fourth and last estimator used in the linear panel data models is the between estimator. This is a different way of estimating the $\beta$, which only exploits differences between households. The formula for this estimation is given in equation 4. In this equation, $\bar{y}_i$, $\bar{x}_i$ and $\bar{\epsilon}_i$ are averaged over the number of time steps $T$. As with the FE model, it is estimated because the results can provide extra insights about the data. However, the expectation is not that this model outperforms the RE model, because it uses less detailed information.

$$\bar{y}_i = \alpha + \bar{x}_i'\beta + \bar{\epsilon}_i \tag{4}$$

**Optimising linear panel data models**

When the most promising linear panel data model is selected, the optimal combination of predicting variables needs to be found. The selection of predictors is explained in this section.

To select the most appropriate set of predictors, forward selection is used. The base model consists of the predictors degree days, surface, energy label and building year, which are expected to be the most important predictors based on literature. In addition, preliminary research found that they all have significant parameters. All other predictors are individually added to this base model, to see if the Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC) improve. All predictors that improved both the AIC and BIC (compared to the base model) are included in the second part of the predictor selection.

In the second part these predictors are added one by one until there is a full model containing all these predictors. All models from the base model to the full model are compared on the AIC, BIC, $R^2$ and parameter estimates. These two steps are executed for data set 1 and 2. The most promising models are compared using out-of-sample estimation.

The formulas for the AIC and BIC are given in equation 5 and 6. In these formulas $k$ is the number of predictors, $n$ is the number of observations and $\hat{L}$ is the likelihood of the regression. The quality of a model always improves (in terms of $R^2$ and (log-)likelihood) when predictors are added. The AIC and BIC therefore balance the quality and the number of predictors, with the goal to minimise the AIC and BIC-values. If the quality barely improves when adding a predictor, the value of the information criteria will go up. This indicates that the true quality of the model has decreased.

$$AIC = 2k - 2\ln(\hat{L}) \tag{5}$$

$$BIC = k\ln(n) - 2\ln(\hat{L}) \tag{6}$$

As explained in the Data chapter, two different data sets are compared. The linear panel models are optimised for both these data sets. A common practice in econometrics is to fit a model on a training sample and then see how it predicts out-of-sample values (which are values from the validation or test set in this study). Because the correct values of the endogenous variable are known, the mean squared prediction error (MSPE, equation 7) and mean absolute percentage prediction error (MAPPE, equation 8) can be computed for these out-of-sample predictions. In these equations, $\hat{y_{it}}$ is the predicted value for $y_{it}$. To see which data set is better to use for prediction, it is common to look at the MSPE and MAPPE, because they show the performance on unseen data. These statistics can be also computed for the training set (they are then just called mean squared error, MSE, and mean absolute percentage error, MAPE). However, as forecasting is the goal of this study, the MSPE and the MAPPE are more important.

$$MSPE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \hat{y_{it}})^2 \tag{7}$$

$$MAPPE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left|\frac{y_{it} - \hat{y_{it}}}{y_{it}}\right| \tag{8}$$

As the some of the models include interaction terms, it is useful to have a closer look at the effects of different predictors on the space heating demand as well. Because of the interaction and squared terms, the coefficient of a predictor $x$ on the space heating demand, $\frac{dx}{dDemand}$, is not just a number, but a function of (mostly other) predictors. The parameters are compared among different models, and are investigated to find out what the effect of a predictor is on the endogenous variable. If a model has very strange coefficients, this could be reason to investigate the model further (and ultimately not to use it).

### 4.2.2 Functional coefficient model

The functional coefficient model considers the coefficients for predictors not constant (like $\beta$ in the RE-model), but varying. This coefficient varies depending on the value of the so-called *varying variable*. However, the relation between the coefficient and the varying variable does not come from a certain distribution; it is non-parametric. Equation 9 gives the expectation of the endogenous variable, given the values of the varying variable ($u$) and the values of other predictors (these are in matrix $\mathbf{X}$). In this equation, the function $a_j(\mathbf{u})$ determines the value of the coefficient for variable $j$ for a certain value of $u$, and $p$ is the number of predictors in $X$.

$$E(Y|U = u, X = x) = \sum_{j=1}^{p} a_j(\mathbf{u})x_j \tag{9}$$

In this research, an adapted version of the functional coefficient model of Cai et al. (2000) is used. The difference is that the predictors in $\mathbf{U}$ and $\mathbf{X}$ are exogenous predictors and not auto-regressive components as was the case in the research of Cai et al. (2000). In this research, the predictor in $\mathbf{U}$ is thus called the varying variable, and the predictors in $\mathbf{X}$ are simply called the predictors. Figure 12 gives an example of how the coefficient of surface depends on the value of degree days (which is the varying variable here). In this very basic model, only floor surface is included as a predictor in $\mathbf{X}$. As a result, the expected space heating demand can be computed using only this figure and the values of surface and degree days for a household in a certain year. Take for example a house with a floor surface of 100 squared metres, and a year with 1800 degree days. The coefficient of surface for 1800 degree days is approximately 0.22 (this is the value of $a_j(\mathbf{u})$ in equation 9). This value is then multiplied with the value of the predictor and thus. the expected space heating demand for this household is approximately $0.22 * 100 = 22$ GJ.

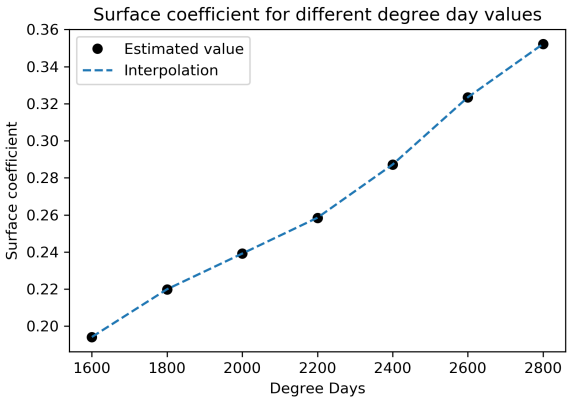


Figure 12: The coefficient of surface increases for higher values of degree days. The figure shows the estimated values (blue dots) and a linear and a cubic interpolation.

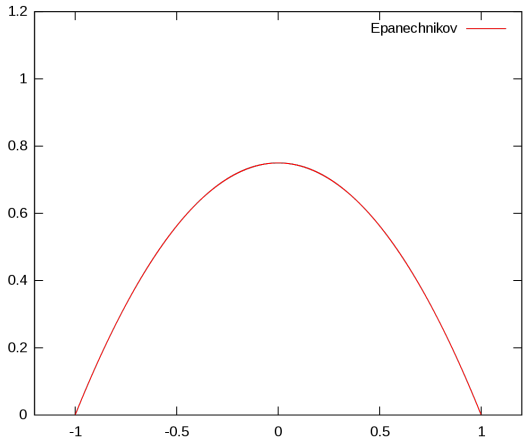


Figure 13: The Epanechnikov kernel. The input is $U_{it} - u_0$, and hence values of $U_{it}$ that are close to $u_0$ obtain higher weights (Amberg, 2008).

### Estimation

The model coefficients are estimated for different values of $u$, naturally spanning the whole range of $u$ in the historical and future data. After these values are estimated, these values are then fitted into a function. The input for the model is $\{\{\mathbf{U}_{it}, \mathbf{X}_{it}, Y_{it}\}_{i=1}^{n}\}_{t=1}^{T}\}$, where $\mathbf{X}_{it} = (X_{it1}, .., X_{itp})^T$, i.e. it contains values for all different predictors 1 to $p$ for observation $i$. The matrix $\mathbf{U}$ can contain multiple variables, but usually contains only one, in which

case it is simply a vector (of length $nT$). The value of $a_j$ at $u_0$ can be approximated by $a_j(u) \approx a_j + b_j(u - u_0)$. Keep in mind here that the coefficient to estimate is in fact the derivative of predictor $j$ to endogenous variable $y$ at value $u_0$. The local linear estimator $\hat{a}_j(u_0)$ can thus be estimated by minimising the sum of weighted squares as in equation 10 over $(a_j, b_j)$. In equation 10, $T$ is the number of time steps, $n$ is the number of observations, $p$ is the number of predictors in $\mathbf{X}$, and $K_h(\cdot) = \frac{1}{h}K(\cdot/h)$. $K(\cdot)$ is a kernel function on $\mathbb{R}^1$ and $h > 0$ is a bandwidth. In this research, the Epanechnikov kernel is used: $K(u) = \frac{3}{4}(1 - u^2)$. The kernel function gives high weights to values for $U_{it}$ that are close to $u_0$, such that values close to $u_0$ have more impact on value of $a_j$ at $u_0$ (see also figure 13).

$$\sum_{t=1}^{T}\sum_{i=1}^{n}\Big[Y_{it} - \sum_{j=1}^{p}\big(a_j + b_j(U_{it} - u_0)\big)X_{itj}\Big]^2 K_h(U_{it} - u_0) \tag{10}$$

From least squares theory, it follows that $\hat{a}_j(u_0)$ can be estimated as in equation 11.

$$\hat{a}_j(u_0) = \sum_{t=1}^{T}\sum_{i=1}^{n} F_{nT,j}(U_{it} - u_0, X_{it})Y_{it} \tag{11}$$

$F_{nT,j}$ in equation 11 is defined as in equation 12.

$$F_{nT,j}(u, x) = e_{j,2p}^T(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}})^{-1}\begin{bmatrix} x \\ ux \end{bmatrix} K_h(u) \tag{12}$$

In equation 12, $e_{j,2p}^T$ is a $2p \times 1$ vector with zeroes and a 1 at its $j$-th entry. As a result, it 'selects' the $j$-th row from $(\tilde{\mathbf{X}}^T\mathbf{W}\tilde{\mathbf{X}})^{-1}$. $\tilde{\mathbf{X}}$ is an $nT \times 2p$ matrix, where row $it$ is $(X_{it}^T, X_{it}^T(U_{it} - u_0))$. The matrix $\mathbf{W}$ is a diagonal matrix of size $nT \times nT$ with $(K_h(U_{11} - u_0), K_h(U_{12} - u_0), ..., K_h(U_{1T} - u_0), K_h(U_{21} - u_0), ..., K_h(U_{nT} - u_0))$ on its diagonal.

The optimal bandwidth $h$ is estimated by experiment. The functions $a_j$ follow from the input value of $h$ using the equations 10 to 12. Having obtained these results, the model quality is determined and compared for different values of $h$ (using the MSE, MAPE, MSPE and MAPPE).

### Optimising the functional coefficient model

To optimise the FCM, first the most appropriate varying variable is determined. As stated, the expectation is that degree days performs well as varying variable. In addition, floor surface is tested as varying variable, because a similar argument could be made for this predictor: if a house would have zero surface, there is nothing to heat and thus the effect of all other predictors is expected to be zero. On the other hand, when a house has a large floor surface, the impact of for example better insulation could be more important than for houses with a

small surface. However, the link between floor surface and the parameter for other predictors is likely to be less direct than the link between degree days and the parameter for other predictors. A cause of this is for example that not all areas of a house a heated to the same extent (e.g. garages and cellars). The expectation is thus, for both variables as varying variable, that the effect of another predictor increases for larger values of the varying variable. This would however not per se mean the parameter *value* increases: for a variable with a negative effect, a stronger, more negative effect is expected for a higher value of the varying variable. The models with the different varying variable are compared using the MSPE and MAPPE.

Having determined the most appropriate varying variable, predictor selection is executed in a similar way as for the linear panel models, but using MSPE and MAPPE to evaluate models. Naturallly, results from the linear panel models are taken into account, to avoid unnecessary work. Interaction effects and quadratic terms are not included. The two data sets are compared using the MSPE and MAPPE as well. Furthermore, the parameter values for different values of the varying variable are investigated. This should give more insight in the structure of the data and helps to determine if the method functions as expected.

To evaluate the coefficient estimates for predictor $j$ at $u = u_0$, $(a_j(u_0))$, bootstrapping is used. Bootstrapping, a concept developed by Efron (1979), is a resampling method to gain insight in the variability of an estimate. From an available sample, a (sub)set is used, where one observation can be used multiple times. The data points are thus selected one-by-one with replacement. "The bootstrap sample is the same size as the original data set. As a result, some samples will be represented multiple times in the bootstrap sample while others will not be selected at all." (Kuhn, Johnson, et al., 2013). In this study, thirty subsets from the training data set are used, in order to be able to fairly compare the estimates with the estimate of the model that performs best. As stated earlier, the FCM optimisation is done using the training set which contains 60% of the data (45548 households).

### 4.2.3 Recurrent Neural Network

The recurrent neural network (RNN) is a specific form of an artificial neural network (ANN). This neural network is used to estimate a non-linear auto-regressive model with exogenous regressors (NARX), as in Koschwitz et al. (2020).

Artificial neural networks are black-box non-parametric data analytics methods, that usually perform well in prediction. More specifically, Wei et al. (2018) found that ANNs perform well compared to other predictive methods in their review of data-driven approaches for build-

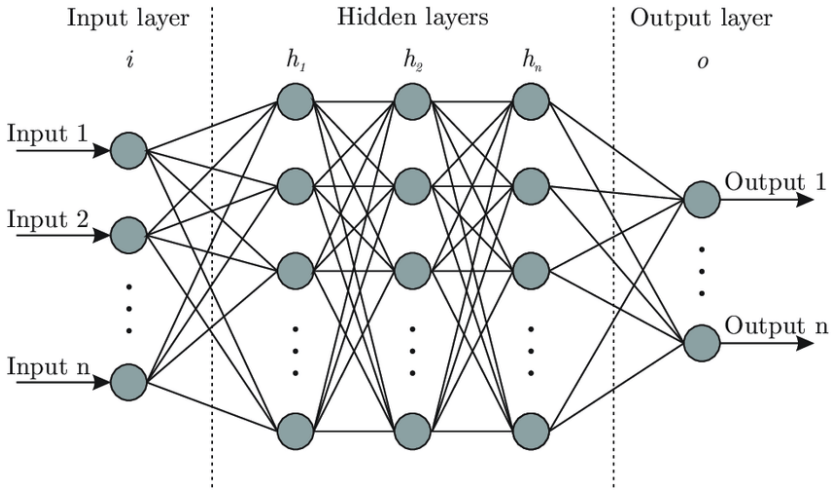


Figure 14: The general structure of an artificial neural network. Image from Bre et al. (2017).

ing energy consumption. The method is inspired by biological neural networks. The idea is to connect (a large amount of) nodes which exchange information to solve a problem. They usually have an architecture as depicted in figure 14. It has an input layer, consisting of the predicting variables, a number of hidden layers with a number of nodes per layer and an output layer. Every node is a function of the values in the layer before, i.e. node $h_{11}$ (node 1 in hidden layer 1) is some function of all values in the input layer, $f(x_1, x_2, ..., x_n)$, with $n$ the number of predictors here. From the last hidden layer, the output is computed, where again the output is some function of all values in the last hidden layer: $f(h_{m1}, h_{m2}, ..., h_{mo})$. Here $m$ is the number of hidden layers and $o$ is the number of nodes in the last hidden layer. Artificial neural networks can capture difficult non-linear structures, which is why they are commonly used for difficult tasks, such as classifying images or texts.

An RNN is an ANN specifically designed for time-series modelling, and thus includes a dynamic effect. The architecture of an RNN is depicted in figure 15. In this figure every green square is a set of hidden layers, that uses not only input from predicting variables at $t$, namely $x_t$ (in blue), but also information from the set of hidden layers from the former time step $t-1$. In this way it can capture dynamic effects. Every set of hidden layers does the same computations, i.e. they perform the same *recurrent* task for every time step. Because every set of hidden layers performs the same task, parameters for only one set of hidden layers have to be estimated. The estimated output values are $h_t$ (in purple).

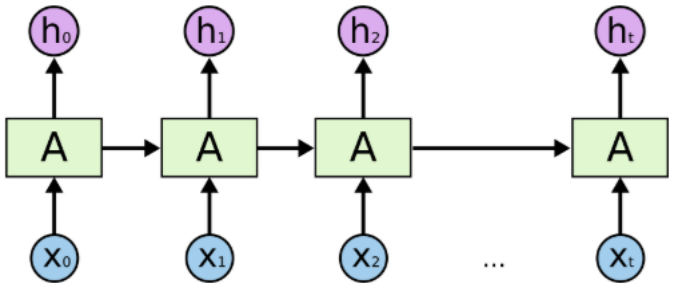


Figure 15: A structural depiction of a recurrent neural network, retrieved from Olah (2015).

**Types of RNNs**

In this section, a basic explanation of 'regular' RNN and the long term short memory (LSTM) is given. Both methods are applied in this research.

A regular RNN can have multiple layers and multiple nodes in one RNN *cell*. In figure 15, a green square depicts an RNN cell. As yet stated, every RNN cell has the same structure in terms of layers and nodes and it thus actually is an artificial neural network in itself (as in figure 14). The inputs are the exogenous variables and the input from the former time step, and the output is the endogenous variable and the input for the next time step. The functions connecting nodes (this includes input and output nodes) in an RNN are called activation functions. In an RNN cell, the tanh function is the most commonly used activation function. (This is also the default in the Keras Python library (Chollet et al., 2015)), which is used to fit the neural networks in this research). An advantage of this activation function is that it always returns an output between -1 and 1. This is important because normalised data improves the learning speed of a neural network (Rafiq, Bugmann, & Easterbrook, 2001). The fitting of the neural network in the end consists of finding the optimal weights for all activation functions and biases. This is done iteratively over a large solution space (the size of course depends on the number of layers, nodes, input and output variables).

The long term short memory (LSTM) model is mainly used for complex data with very long sequences. The LSTM is an extension of the regular RNN that is able to find long term relations more easily. A structural depiction of a basic LSTM cell is given in figure 16. Within an LSTM cell, information flows are controlled by the forget gate ($f_t$ in figure 16) and the input gate ($i_t$ in figure 16). The forget gate is actually a function that determines what information from the last cell state ($C_{t-1}$) can be forgotten, based on new inputs $x_t$ and the output value of the former cell $h_{t-1}$. Usually this is a sigmoid function, symbolised by the $\sigma$. The input gate is a function that determines how the input values $x_t$, and the former

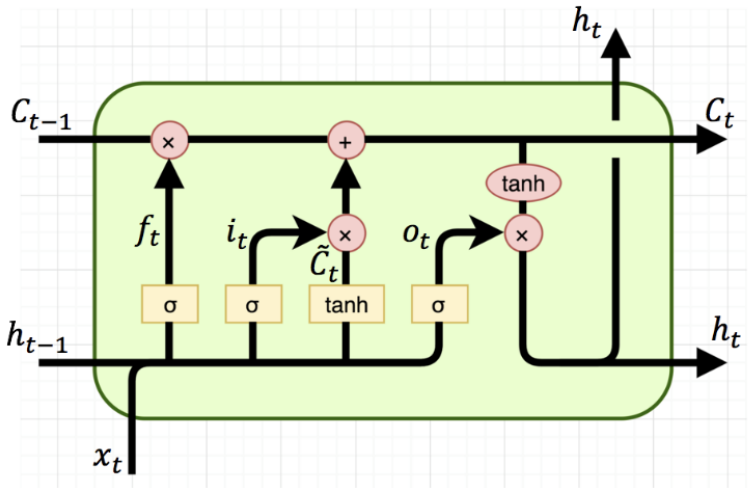


Figure 16: The architecture of a basic LSTM cell, retrieved from Mani (2019).

output value $h_{t-1}$ are updated. In addition, a tanh layer transforms again $x_t$ and $h_{t-1}$ to create candidate cell values $\tilde{C}_t$. Subsequently, the candidate cell values and the results from the input gate layer are multiplied: $i_t * \tilde{C}_t$. The current cell state $C_t$ is now determined, by forgetting what should be forgotten (using $f_t$) and adding $i_t * \tilde{C}_t$. Then, conclusively, the new output $h_t$ is computed, after once again transforming the cell state by a tanh layer and multiplying this by $o_t$, which is a sigmoid function of again the predictors at $t$ ($x_t$) and the former output $h_{t-1}$. Several extensions to this basic LSTM cell are possible, but these are not used in this research and thus not discussed here. The Keras Python library is used to create the LSTM (Chollet et al., 2015) with the default LSTM activation functions as stated in this paragraph.

**Estimating (hyper)parameters**

Hyperparameters are input parameters for a model. They are not estimated by the model but a given, and it is thus important to experiment with hyperparameters to find the best model settings (this is called 'tuning' hyperparameters). The parameters for all activation functions and biases are estimated when fitting the model on the data, and differ for different hyperparameter settings (and also for different runs with the same hyperparameters).

Clearly, the solution space for an RNN is very large, as all activation functions and gates need parameters for many variables. The optimisation algorithm to find the optimum value is Adam, developed by Kingma and Ba (2014), which they describe as "computationally efficient, has little memory requirement, invariant to diagonal rescaling of gradients, and is well

suited for problems that are large in terms of data/parameters". The loss function is the mean squared error (MSE), to be able to compare this method fairly with the other methods (which optimise on the MSE as well). The number of epochs is the number of iterations over the complete data set. This is the first hyperparameter to tune in practice. Generally, the number of epochs is optimised looking at the MSE of the validation set. However, as the run times increase with more epochs, a balance needs to be found between more epochs and a better validation MSE.

A second hyperparameter is the batch size. The model updates its values after every batch. A batch of 10 contains all the data for 10 households. The number of layers and the number of nodes per layer are other hyperparameters that need to be tuned. When the most suitable hyperparameters are determined, the model with these hyperparameters is fit on the training set and the parameters are determined. Because the solution space is very large and often non-convex, different runs to optimise the model result in different model parameters. As a consequence, it is necessary to execute multiple runs (usually 10 to 20) to fairly compare different model settings. The set of predictors is determined based on the results from the other two methods. However, some experimenting is possible if there was not a clear optimal set of input variables found by the first two methods. Different hyperparameter settings are compared using the MSPE and MAPPE.

## 4.3 Method evaluation

The methods are evaluated using various quality measures, but the most important are again the MSPE and MAPPE. Optimism is used to check for overfitting: $O = MSPE - MSE$. High values of optimism are a sign of overfitting, because the model describes the training data much better than the validation data. In the end, performance quality is measured combining the results of the described performance indicators, including the MSE and MAPE. In case of ambiguity, generally the MSPE is leading. The methods are compared using the training and validation set (combined) as input, and the test set as out-of-sample set.

To check the validity of the method evaluation, the models are trained on different sets (also of different sizes). The basic training set is the combination of the training and validation set, containing 80% of the data, as is described in section 3.3.4. However, the best models per method are also trained on training sets containing 85, 90 and 95% of the data. Subsequently these models then predict the out-of-sample data containing 15, 10 and 5% of the data set respectively. This is also a way of evaluating the split in data.

### 4.4 Forecast combination

After the three methods are compared, the predictions are combined to see if this improves prediction quality. As became clear from the literature, using equal weights usually performs better than trying to find optimal weights. As a consequence, equal weights are used throughout this study. If the combined prediction performs better than the individual predictions of all methods (in predicting the test set), the predictions are combined for forecasting as well.

### 4.5 Hypotheses

In this section, hypotheses are established for all research questions.

1. What is the effect of degree days, a building's energy label and other building characteristics on space heating demand?

   (a) The effect of degree days on space heating demand is expected to be positive. Preliminary research has shown that up to almost 85% of the variance of the total heating demand can be explained with degree days only. It is thus expected to be the most important predictor for space heating demand.

   (b) A significant negative effect of insulation on space heating demand is expected (i.e. label A has a lower demand than label G). Existing research has shown that theoretical gains of better insulation are higher than practical gains (Majcen et al., 2013). Still, the effect is expected to be measurable (and significant), as it was in the research of Majcen et al. (2013).

   (c) Some building characteristics are expected to have effects on the space heating demand - building year and building typology in particular. For building year, newer buildings are expected to have a lower space heating demand (thus a negative effect is expected). The building year might, however, be related to insulation quality - and thus its effect could thus not be significant. Building typologies with more shared walls are expected to have lower space heating demand. The floor surface of a building is expected to have a positive effect on space heating demand as well.

2. What is the expected space heating demand for a household from 2020 until 2050 and what is the variability of this prediction?

   (a) A decrease of about 1,5 % per year is expected (as found by Segers et al. (2019)), resulting in a decrease in 2050 between 22 and 52% (similar to results found by Andrić et al. (2016)). The variability is expected to increase with the time and thus be higher in 2050 than in 2030 e.g.

3. What is the most appropriate econometric method to model the relationship between degree days, a building's energy label and certain building characteristics on the one hand, and space heating demand on the other hand?

   (a) About linear panel data models: The random effects (RE) model is expected to perform best within the linear panel data models, as it is able to include occupant behaviour and time-invariant regressors. Hence, the expectation is that the random effects model outperforms the fixed effects model as well (because of the inclusion of time-invariant regressors). Although it is likely that the most important predictor (degree days) has a linear relationship without the output variable, the RE model is expected to perform worse than other methods because of the limited flexibility.

   (b) About the functional coefficient model (FCM): This model is expected to add prediction quality to the RE model. The FCM is more flexibile due to the nonlinearity in the parameters. In addition, there is a reason to expect that the parameters for some predictors depend on the number of degree days. When there are no degree days for example, it is unlikely that any other predictor causes space heating demand. On the other hand, on very cold days, it is likely that insulation quality results in a higher reduction of space heating demand than on days with temperatures just below the degree day threshold.

   (c) About the recurrent neural network: The RNN is expected to perform best on prediction quality. However, due to the black-box structure, it does not improve understanding of the causal relations in the system. The LSTM, although promising in many fields, might be sensitive for overfitting. The data are not extremely complex and do not contain very long sequences. That is why the performance of a regular RNN network is expected to be better.

   (d) To conclude, the RNN is expected to perform best in describing the relationship between the variables and in prediction. It is expected that this method could outperform the other methods by some distance because of its flexibility. However, the RE model and the FCM are expected to contribute to interpretability and as a consequence improve the understanding of the relations in the system.

# 5 Results

In this section the results of the research are presented. Firstly, the main results of the optimisation of the three methods is dealt with shortly (section 5.1). After the method optimisation, the three methods are compared in section 5.2. At last, the predictions for future space heating demand are presented (section 5.3).

## 5.1 Method optimisation

In this section, the results from the method optimisation are shortly presented per method. For an extensive explanation of the optimisation per method, readers are referred to the appendices B.1 for linear panel data models, B.2 for the Functional Coefficient Model and B.3 for the Recurrent Neural Network.

### 5.1.1 Linear panel models

From the four tested models, the random effects model performs best, having the highest $R^2$-values. This model performs much better in predicting training set than out-of-sample data. This is caused by the household specific intercept, which cannot be included in out-of-sample prediction. Training the model on data set 1 results in better prediction than training on data set 2.

The predictors included in the model are degree days, floor surface, building year, energy label, house type and the interaction terms surface · degree days and building year · year. The models are not only compared on the validation set, but also on parameter values and performance in forecasting.

All parameters mentioned in the hypotheses had a significant effect with the expected sign. This means that degree days, floor surface and energy label have a positive effect, and building year has a negative effect. For the house types, indeed houses surrounded by more open space have higher space heating demands.

### 5.1.2 Functional Coefficient Model

The varying variable is degree days. Models with these varying variable perform well, while models with floor surface as varying variable show some irregularities. The predictors included are same as in RE model: degree days, floor surface, building year, energy label and house type. No interaction effects are included and neither is an intercept (this causes overspecification in combination with including degree days as varying variable and predictor).

Training on data set 1 again resulted in models that predicted better. As a consequence, this data set is used from this point on in the research. Excluding data set 2, naturally, decreases run times by half. This is useful especially because neural networks are more complex and thus require longer run times.

The parameter values increase when the varying variable increases, as expected. The parameter values are generally stable using bootstrapping (i.e. bootstrap results have little dispersion). For the house types the standard deviations are larger sometimes. However, different parameter estimates do barely affect prediction quality, because all 30 bootstrap samples resulted in MSPEs between 332.8 and 333. Therefore, it seems that the dispersion only occurred for less important predictors.

### 5.1.3 Recurrent Neural Network

The long short term memory (LSTM) network performs badly. The MSPE increases strongly after few epochs (see figure 41 in appendix B.3), and this is the case for all hyperparameter settings tested. In addition, performance after very few epochs is bad in terms of prediction quality. Overfitting seems to occur, based on the increase of the loss function in the validation set after a few epochs. In addition, very simple models perform better than more complicate ones (i.e. models with fewer nodes and fewer layers have lower loss values). This is an indication of overfitting as well.

The regular recurrent neural network (RNN) does function well. The (squared) loss values decrease strongly in the first 20 epochs, and then slowly plateau after about 100 epochs (see figure 42 in appendix B.3). The variables included are again degree days, floor surface, building year, energy label and house type. The hyperparameter testing found that the model trained in 100 epochs, with batch size 96, 2 layers and 7 nodes per layer performed best (on the MSPE). This RNN is thus used in the remainder of this research.

## 5.2 Method comparison

To make a fair comparison between the three methods, their performance is tested on the (so far unseen) test set. They are trained on the data set containing the training and validation set. As a recap, table 3 gives a short description of the three best models per method.

Table 4 shows the performance of the methods on the most important quality measures, for different sizes of the training set. In general, it can be stated that none of the methods performs very well in predicting the yearly space heating demand of individual households. Considering that the yearly mean demands are between 21 and 37 GJ (see table 7 in appendix

Table 3: Explanation of the settings and predictors included for the best performing model per method.

| Method | Explanation |
|---|---|
| RE | Includes the predictors degree days, building year, energy label, floor surface and house type. Also included are the interaction term of floor surface and degree days and the interaction term of building year and year. |
| FCM | The varying variable is degree days, and the predictors are energy label, floor surface, house type, building year and degree days. |
| RNN | This is a neural network of regular RNN cells that has 2 layers and 7 nodes per layer. It is fitted with 100 epochs and batch size 96. It includes the variables degree days, building year, floor surface, energy label and house type. |

A.3), mean squared prediction errors of above 300 are high. In addition, MAPPEs above 1 can be considered high in general as well. Possible causes could be occupant behaviour, the fact that households might move relatively often and dealing with a very noisy data set.

The RNN performs best on the main quality indicator, the mean squared prediction error, for all test sets. Relevant to note is that the RNN performance on the MSPE is fairly stable for different runs. Having executed 10 runs (for all set sizes), the range of the MSPEs is between 2 and 5 wide. The RNN method however scores worst on the mean average percentage prediction error. The other three methods (including the combined one) perform much better on the MAPPE. The combined estimation performs almost as well as the RNN on the MSPE.

It was expected that the RNN method would perform best for prediction (at least compared to RE and FCM), and it does perform best on the MSPE, which is the main performance indicator. However, the improvements compared to the other methods are very limited. For all data sets, the maximum improvement compared to the worst performer is 2.5%. This indicates that the data are not extremely complex in terms of relations between variables. Relative simple structures in the data are in general the likely cause of the similar performance of the methods. Combining this insight with the fact that the prediction performance is not so well, leads to the conclusion that the data might contain a lot of noise.

In general, the MSPE decreases when using larger training sets, with the exception of the training set that contains 90% of the data. For the MAPPE this is not the case however. For all methods, the MAPPE is largest for the training set containing 95% of the data. The latter, naturally, is an unexpected result.

The most important result however is the performance of the combined prediction. As it

Table 4: The main quality measures for the most promising model per method. The training and validation set were used for training (together) and the test set for testing. The top two rows show the performance indicator (MSPE or MAPPE) and the relative set size of the training set.

| | MSPE | | | | MAPPE | | | |
|---|---|---|---|---|---|---|---|---|
| | 80 | 85 | 90 | 95 | 80 | 85 | 90 | 95 |
| RE | 339.8 | 334.6 | 338.4 | 312.3 | 1.43 | 1.41 | 1.66 | 1.77 |
| FCM | 339.4 | 334.5 | 338.3 | 312.2 | 1.41 | 1.40 | 1.66 | 1.78 |
| RNN | 335.7 | 327.5 | 329.9 | 306.1 | 1.79 | 1.73 | 1.95 | 2.05 |
| Combined | 336.4 | 329.0 | 332.5 | 307.3 | 1.45 | 1.41 | 1.65 | 1.77 |

performs well, it is used for forecasting. The reason for this is twofold. Firstly, it performs best looking at the MSPE and MAPPE. On the MSPE is scores almost as good as the RNN, and on the MAPPE is scores much better. Secondly, it is a robust way of forecasting that has proven to be effective in many other studies.

It can be concluded that different training/test set splits give very similar results in terms of the order of method performance. For the MSPE, the RNN always scores best, the combined forecast is a close second and RE and FCM perform similar but clearly worse than the others. For the MAPPE, RNN obviously performs less than the other three options, which have similar scores. It thus indicates that the chosen training/test splits are valid. In addition, it shows that the performances and the order hereof are valid too.

The parameters for the predictors are similar in the RE model and the FCM. Tables 16 and 17 in appendix C show all parameter values. However, because of the interaction terms in the RE model, they cannot be fully compared. As earlier explained, it is possible to compute the relative effects of a predictor given the values of the other predictors. However, when you do this for all the predictors and you consider these relative effects the same as parameters, you count the interaction effects double. For the FCM, the parameters are taken at a value of 2200 for degree days. Nonetheless, when comparing the relative effects, most are very similar. The relative effect of floor surface is between 0.20 and 0.24, for building year between -0.061 and -0.064, and for energy label between 1.22 and 1.33. The relative effect for degree days is around 0.057 in the FCM and only 0.013 in the RE models . The RE model, however, does have an intercept of about 120 that is added (which the FCM does not have). The difference in degree days compensates about 95 of this 120 for the average value of degree days. In addition, most other parameters have a slightly higher coefficient for FCM than for the RE model. The parameters for the house types are very similar as well. In addition, all relative effects have the expected signs.

## 5.3 Future space heating demand

In this section the forecasting results are presented. First forecasts per method are computed by feeding future data sets to the models of the three methods. Then the combined forecasts are made by taking the average value of the forecasts of the three methods. The future data sets are based on the climate scenarios and energy label scenarios as described section 3.6. As for both degree days and energy labels there are 4 possible scenarios, this results in 16 main scenarios. In this section the results for the combined forecasts are presented at first. Hereafter, the results for the different methods are compared. All figures in this section show the historical values available from the data used in this study, and a trend line. These data place the forecasts in some perspective. It is, however, important to note that very few historical data points are available, and it is thus incorrect to draw (strong) conclusions from these historical data points and the accompanying trend line. Nor is it correct to draw conclusions about the forecasts being either valid or invalid based on this depiction of history.

### 5.3.1 Combined forecasts

The averages of all forecasts for 2020, 2030, 2040 and 2050 are 28.6, 26.3, 24.2 and 22.4 GJ respectively. The total relative decline in demand is 21.6%. This is an average yearly decline of 0.8%. The decline decreases slowly in absolute terms, but the relative decline seems more stable with 7.8% in the first decade, 8.2% in the second and 7.4% in the last decade. Figure 17 shows a boxplot of the combined forecasts for all 480 scenarios (30 weather scenarios per year, 4 climate scenarios and 4 energy label scenarios) for 2020, 2030, 2040 and 2050. The ranges are clearly wide. The width between the 5th and 95th percentile (these are shown in table 5) increase slowly over time from 11.6 GJ in 2020 to 12.5 GJ in 2050. From the assumption that every scenario is as likely as another, one could consider the range between the 5th and the 95th percentile a 90% confidence interval for the predictions.

The combined forecasts for the 16 main scenarios are shown in figure 18. It shows a steady decline of expected space heating demand for all scenarios. The smallest relative decrease in the period 2020-2050 is 15.4%, corresponding to an annual average decrease of 0.56%. This smallest decrease occurs in the GL tide scenario. The largest relative decrease occurs in the WH circles scenario. Between 2020 and 2050 the heat demand lowers with 28.0%, which is 1.09% annually. As a result, the space heating demand is highest in the GL tide scenario in 2050, 25.5 GJ, and lowest in the WH circles scenario, 19.6 GJ. Taking into account the 90% scenarios however, results in a range from 15.2 to 30.9 GJ.

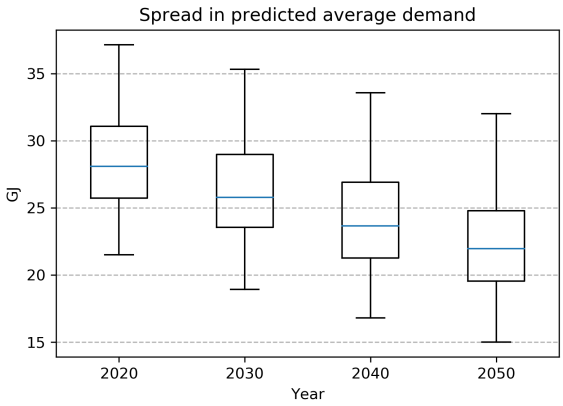


Figure 17: Boxplot of the combined forecast for all scenarios in 2020, 2030, 2040 and 2050.

Table 5: The 95th, 50th (i.e. the median) and 5th percentile corresponding to the boxplot in figure 17.

| Percentile | 2020 | 2030 | 2040 | 2050 |
|---|---|---|---|---|
| 95 | 35.0 | 33.0 | 31.1 | 29.7 |
| 50 | 28.1 | 25.8 | 23.6 | 21.9 |
| 5 | 23.4 | 21.2 | 19.1 | 17.2 |

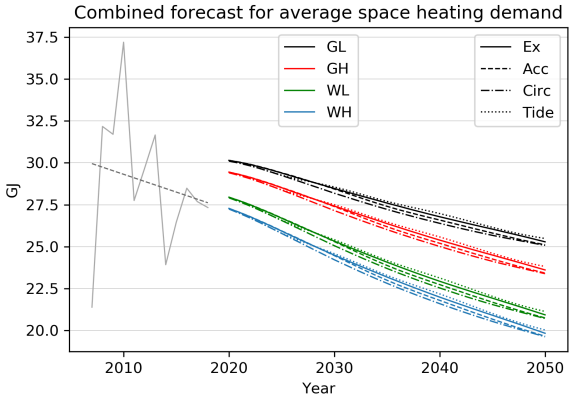


Figure 18: The combined forecast for averaged space heating demand in the sixteen main scenarios.

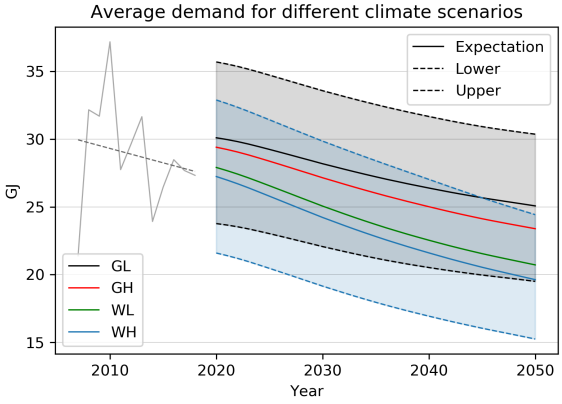


Figure 19: The forecast for different climate scenarios (for the circles energy label scenario). It shows that the differences between the main climate scenarios increase over time, and that the variability within a climate scenario is very large.

Figure 18 shows that the variability in scenarios is mainly caused by the climate scenarios. The energy label scenarios cause differences of around 0.41 GJ in 2050. The differences in these scenarios peak in around 2040, which is plausible given the setup of the energy label scenarios. However, even at this peak, the maximum differences are always lower than 0.6 GJ (looking at the main scenarios only). The climate scenarios, on the other hand, cause maximum differences of about 5.4 GJ in 2050 (between WH and GL scenarios). In addition, this is only the spread in the main scenarios. In these main scenarios, the possible spread within climate scenarios is already excluded. If the 90% scenarios are included as climate scenarios, differences between these scenarios can increase up to 15 GJ. Figure 19 shows that the differences in the main scenarios (solid lines) slowly increase. Adding the 90% interval borders (dotted lines), increases the range of possibilities strongly.

### 5.3.2 Method comparison

All methods show the same pattern of a steady and relatively linear decrease. However, the steepness of the decrease and starting levels differ. Figure 20 shows this for the GL accelerated policy and WH circles scenario. The RNN forecasts higher heat demand than the other methods, and the RE method forecasts the lowest heat demand. The RNN also forecasts very small differences in energy scenarios. Figure 21 shows that the different energy label scenarios for RNN (in green) are closer to each other than the different energy label scenarios for the other methods. This is shown here for the WL climate scenarios only, but this is the case for all climate scenarios. The forecasts for different energy label scenarios for the RE and FCM display a larger influence of the energy label scenarios. The figure also shows a slight change in the yearly decrease for the RNN (the slope becomes less steep around 2035), which is not observable for the RE method and the FCM. A similar pattern is seen in the RNN forecasts for the WH climate scenario.

For the RE model, the main scenarios range from 17.5 to 23.4 GJ in 2050. The FCM has higher values, ranging from 19.3 to 25.6 GJ. The RNN had higher forecasts from the beginning and also in 2050 has the highest forecasts, ranging from 22.1 to 27.45 for the main scenarios. The ranges from the RE model and the RNN thus have only a small overlap. The average yearly percentage decrease differs as well between the methods. For RE these are between 0.7 and 1.4%, for FCM between 0.4 and 1.0% and for the RNN between 0.4 and 0.7%. Apart from the fact that the RNN thus forecasts higher values in the coming years, it also predicts a smaller yearly decrease. This explains the large differences in space heating demand between the methods in 2050.

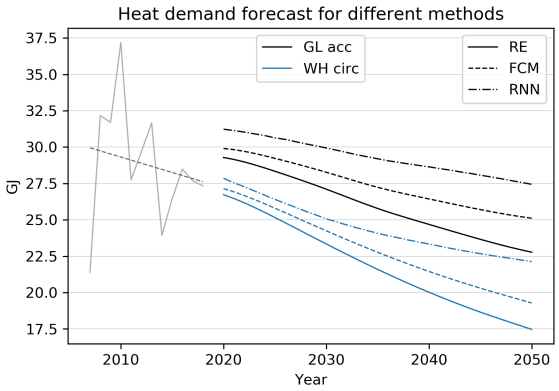


Figure 20: The differences between the three methods for the GL accelerated and WH circles scenario.

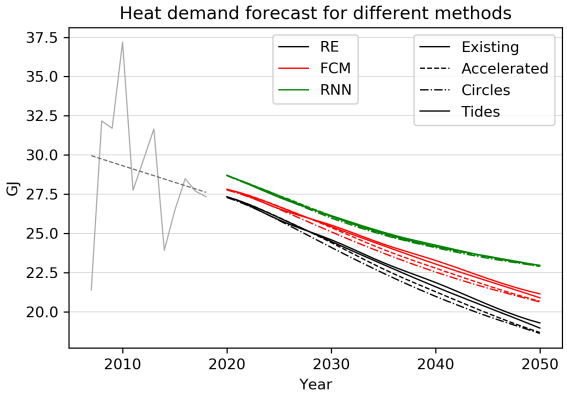


Figure 21: The forecast demand for different methods and different energy label scenarios in the WL climate scenario.

To conclude, the three different methods show generally the same (expected) pattern: a steadily declining demand. The steepness of this decline and the starting point however differ. In addition, methods that forecast higher demand from the start, do not make up for the decline, but differences increase over the years. The differences in forecasts between the methods, increase the value of the forecast combination, especially because the prediction quality for the methods was so similar (see section 5.2).

# 6 Conclusion

In the conclusion the results for the three research questions are presented. Hereafter, shortcomings and remarks regarding the study are discussed.

## 6.1 The effects of the predictors

The effect of degree days on space heating demand is significant and strong. Both the RE model and the FCM find this. The forecasts for all the methods confirm the importance of degree days, as the climate scenarios are of a large impact on the future space heating demand.

The effect of a building's energy label is present but limited. The RE model and the FCM estimate that the average effect of one improvement step of the energy label (for example from C to B) results in around 1.25 GJ less demand per year. The forecasts show that the impact of a quick transformation to an energy efficient housing stock compared to a slow transition is not so big. This is especially the case when comparing this with the impact of the climate scenarios.

For building year, the RE model and the FCM find a parameter of (around) -0.06. This indicates that, for example, if a house would be 50 years older, the space heating demand is expected to be 3 GJ less per year. For floor surface, both models find coefficients of around 0.22. The consequence would be that, on average, a house with 25 squared metres extra is expected to have 5.5 GJ extra yearly space heating demand.

For all the predictors, the sign of the parameter is as expected. In addition, the importance of different predictors is generally in line with the hypotheses. However, the variable floor surface is more important than expected.

## 6.2 Future space heating demand

For the analysed period of 2007 to 2018, an average space heating demand of 28.8 GJ is found. The predicted means for 2020, 2030, 2040 and 2050 are 28.6, 26.3, 24.2 and 22.4 GJ respectively. The insecurity for these predictions is considerably large due to natural fluctuations in weather and insecurity about the future climate. The 90% confidence interval ranges from 17.2 to 29.7 GJ in 2050. A total decline of 21.6% is expected from 2020 to 2050, which is an average yearly decrease of 0.81%.

This 0.81% is much smaller than the hypothesis of a yearly decline of 1.5% and, naturally, the expected decrease over thirty years (21.6%) is much smaller than the hypothesis as well

(36.5%). The decline range found by Andrić et al. (2016) for Lisbon was 22.3 to 52.4% from 2020 to 2050. Again, this study finds a lower decline: for the 16 main scenarios the decline ranges from 15.4 to 28% over the same period.

The variability increases in time between the climate scenarios but not within the climate scenarios. This is caused by the structure of the climate scenarios. Between these scenarios, the variability in terms of degree days increases, but the variability in number of degree days within the scenarios does not increase.

## 6.3 Method performance

In general, the performance of the methods is comparable. The results regarding predicting test sets are relatively small. The RNN performs only slightly better than other methods on the MSPE. It can thus be stated that for the data available, simpler methods are preferred. Here, it is taken into account that the RNN is a more complex model that costs both more time to develop and to run. As the RNN is the only method including a dynamic effect, a dynamic effect is unlikely to be strong. This can be considered an unexpected outcome, because one would expect that a household with a high space heating demand does so every year (and the same for households with a low space heating demand). The RE model and the FCM perform very similar in prediction quality. Although the FCM seemed a very good method for these data, it barely improves prediction quality compared to the RE model. From the fact that all methods perform similar, it can be concluded that the structures in the data are likely to be relatively simple. The more complex methods (FCM and RNN in this study), however, might perform better with more detailed data. When daily or even hourly demand is known, more complex structures might be captured. In addition, for better prediction one could wish for a proxy for occupant behaviour.

## 6.4 Discussion

In this section shortcomings and some relevant remarks regarding this research are presented. In addition, future research on this topic is discussed.

Starting with the data, it is clear that a higher granularity would provide more information. The yearly data is highly aggregated, and causes a sizeable loss of information. On the other hand, for predicting average demands on a yearly basis, this granularity should be sufficient. Still, it is probable that more complex structures could be found with higher granularity.

The data set contains a substantial amount of households (more than 100,000), given the

fact that there are almost 8 million households in the Netherlands (CBS, 2019). The houses that are already connected to a district heating network however, can be considered 'quick-wins', or at least the easier part of the housing stock to connect to a district heating network. It is very uncommon for rural households to be connected to a district heating network, and as a result the largest part of the houses are situated in cities and suburbs. Because of the type of buildings and so-called heat islands, the average heating demand in the Netherlands might be higher than found in this research. On the other hand, there is no obvious reason to assume that the relative predicted decrease in yearly space heating demand would be different for the Dutch housing stock as a whole.

The available data was fairly noisy. Finding heat demand values of over 100,000 GJ suggests mistakes in administration. Naturally, it is unclear for non-extreme values if these were administrative mistakes as well. The energy label is also a difficult proxy. There are only a few moments that an energy label needs to be determined, so improvements in energy efficiency in the meantime remain invisible. In addition, the energy label takes into account a wide range of sustainability issues. The presence of solar panels for example improves the energy label, but does not affect the space heating demand. To solve these issues, the input data for computing the energy label is needed.

A third data-related issue is the lack of information on the level of energy efficiency of the future housing stock. To the author's knowledge, no data on this is available. Furthermore, this is very hard to predict, as such predictions are strongly dependent on political decisions.

A fourth issue to discuss is occupant behaviour. Earlier research found that this is very difficult include in a model and very influential, and this is confirmed in this research. It is the likely cause of the limitation in prediction quality of all the methods. It is recommended to conduct future research on different aspects of occupant behaviour. First, it should be investigated if this indeed is the factor that explains (much of) the noise still present. If this is the case, research on possible proxies for occupant behaviour is needed. A first thought on this is that such a proxy could be a wide range of predictors, from very concrete, like the number of occupants, to more far-fetched, like the political preference (where 'green' voters are expected to have a lower space heating demand). A last remark is that if occupant behaviour indeed has such a strong impact on the space heating demand, this should be taken into account in the science and policy regarding energy efficiency in houses. To give an example, when insulation quality is good, it might be better to start a campaign focusing on behaviour instead of improving insulation even more. Such a campaign could result in a higher reduction of energy demand than the insulation improvements.

With regard to finding the most appropriate econometric method to model and forecast space heating demand, it must be noted that, naturally, it is impossible to test all econometric methods. In this research, three very different methods are tested to get a first idea of in what direction needs to be searched to find the most appropriate econometric method. It has shown that with this type of data (high spatial granularity and low temporal granularity), simpler methods are preferred.

## References


Alfons, A. (2019). *Robust estimation of multivariate location and scatter.*

Amasyali, K., & El-Gohary, N. M. (2018). A review of data-driven building energy consumption prediction studies. *Renewable and Sustainable Energy Reviews*, *81*, 1192–1205.

Amberg, B. (2008). *A graph of the epanechnikov kernel.* Retrieved from https://en.wikipedia.org/wiki/Kernel_(statistics)#/media/File:Kernel_epanechnikov.svg

Andrić, I., Gomes, N., Pina, A., Ferrão, P., Fournier, J., Lacarrière, B., & Le Corre, O. (2016). Modeling the long-term effect of climate change on building heat demand: Case study on a district level. *Energy and Buildings*, *126*, 77–93.

Bates, J. M., & Granger, C. W. (1969). The combination of forecasts. *Journal of the Operational Research Society*, *20*(4), 451–468.

Bre, F., Gimenez, J., & Fachinotti, V. (2017, 11). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings*, *158*. doi: 10.1016/j.enbuild.2017.11.045

Brown, R. H., Clark, D., Corliss, G. F., Nourzad, F., Quinn, T., & Twetten, C. (2012). Forecasting natural gas demand: the role of physical and economic factors. In *32nd annual international symposium on forecasting.*

Cai, Z., Fan, J., & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, *95*(451), 941–956.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications.* Cambridge university press.

Catalina, T., Iordache, V., & Caracaleanu, B. (2013). Multiple regression model for fast prediction of the heating energy demand. *Energy and buildings*, *57*, 302–312.

CBS. (2019). *Huishoudens.* Retrieved from https://opendata.cbs.nl/statline/#/CBS/nl/dataset/71486NED/table?fromstatweb

Changlin, M., & Ning, W. (2001). Functional-coefficient regression model and its estimation. *Applied Mathematics-A Journal of Chinese Universities*, *16*(3), 304–314.

Chollet, F., et al. (2015). *Keras.* https://keras.io.

Claeskens, G., Magnus, J. R., Vasnev, A. L., & Wang, W. (2016). The forecast combination puzzle: A simple theoretical explanation. *International Journal of Forecasting*, *32*(3), 754–762.

Cox, R. A., Drews, M., Rode, C., & Nielsen, S. B. (2015). Simple future weather files for estimating heating and cooling demand. *Building and Environment*, *83*, 104–114.

Dascalaki, E. G., Droutsa, K. G., Balaras, C. A., & Kontoyiannidis, S. (2011). Building typologies as a tool for assessing the energy performance of residential buildings–a case study for the hellenic building stock. *Energy and Buildings*, *43*(12), 3400–3409.

De Groene Amsterdammer. (2018). *Gasverslaafd.* Retrieved from https://www.groene.nl/artikel/gasverslaafd

De Nederlandse Rijksoverheid. (2020). *Hoe lang kan ik nog koken en stoken op gas?* Retrieved from https://www.rijksoverheid.nl/onderwerpen/duurzame-energie/vraag-en-antwoord/hoe-lang-kan-ik-nog-koken-op-gas

Dolinar, M., Vidrih, B., Kajfež-Bogataj, L., & Medved, S. (2010). Predicted changes in energy demands for heating and cooling due to climate change. *Physics and Chemistry of the Earth, Parts A/B/C*, *35*(1-2), 100–106.

Efron, B. (1979). The 1977 rietz lecture. *The annals of Statistics*, *7*(1), 1–26.

Energeia. (2020). *'gas heeft ons veel gebracht, daar hebben we nu last van'.* Retrieved from https://energeia.nl/energeia-artikel/40086909/gas-heeft-ons-veel-gebracht-daar-hebben-we-nu-last-van

European Parliament and Council. (2002, December). *Directive 2002/91/ec of the european parliament and of the council of 16 december 2002 on the energy performance of buildings.*

European Parliament and Council. (2010, May). *Directive 2010/31/eu of the european parliament and of the council.*

Fang, T., & Lahdelma, R. (2016). Evaluation of a multiple linear regression model and sarima model in forecasting heat demand for district heating system. *Applied energy*, *179*, 544–552.

Fekkes, M. (2020). personal communication.

Frank, T. (2005). Climate change impacts on building heating and cooling energy demand in switzerland. *Energy and buildings*, *37*(11), 1175–1185.

Frederiksen, S., & Werner, S. (2013). District heating and cooling (studentlitteratur ab). *Lund, Sweden*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Jylhä, K., Jokisalo, J., Ruosteenoja, K., Pilli-Sihvola, K., Kalamees, T., Seitola, T., ... Drebs, A. (2015). Energy demand for the heating and cooling of residential houses in finland in a changing climate. *Energy and Buildings*, *99*, 104–116.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klein Tank, A., Beersma, J., Bessembinder, J., Van den Hurk, B., & Lenderink, G. (2014). Knmi 14: Klimaatscenario's voor nederland. *KNMI publicatie*.

KNMI. (2020a). *Daggegevens van het weer in nederland - download.* Retrieved from http://projects.knmi.nl/klimatologie/daggegevens/selectie.cgi

KNMI. (2020b). *Knmi klimaatscenarios.* Retrieved from http://climexp.knmi.nl/scenarios_knmi14_form.cgi

Kong, W., Dong, Z. Y., Jia, Y., Hill, D. J., Xu, Y., & Zhang, Y. (2017). Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid*, *10*(1), 841–851.

Koschwitz, D., Spinnräker, E., Frisch, J., & van Treeck, C. (2020). Long-term urban heating load predictions based on optimized retrofit orders: A cross-scenario analysis. *Energy and Buildings*, *208*, 109637.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling* (Vol. 26). Springer.

Labandeira, X., Labeaga, J. M., & López-Otero, X. (2017). A meta-analysis on the price elasticity of energy demand. *Energy Policy*, *102*, 549–568.

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data. john wiley & sons. *New York*.

Lopes, L., Hokoi, S., Miura, H., & Shuhei, K. (2005). Energy efficiency and energy savings in japanese residential buildings—research methodology and surveyed results. *Energy and buildings*, *37*(7), 698–706.

Majcen, D., Itard, L., & Visscher, H. (2013). Theoretical vs. actual energy consumption of labelled dwellings in the netherlands: Discrepancies and policy implications. *Energy policy*, *54*, 125–136.

Mani, K. (2019). *Gru's and lstm's.* Retrieved from https://towardsdatascience.com/grus-and-lstm-s-741709a9b9b1

Mastrucci, A., Baume, O., Stazi, F., & Leopold, U. (2014). Estimating energy savings for the residential building stock of an entire city: A gis-based statistical downscaling approach applied to rotterdam. *Energy and Buildings*, *75*, 358–367.

Naghi, A. A. (2018). *Applied microeconometrics: linear static panel data model.*

Ng, A. (2020). *Gru's and lstm's.* Retrieved from https://www.coursera.org/lecture/deep-neural-network/train-dev-test-sets-cxG1s

Nijpels, E. (2019). *Klimaatakkoord.* Retrieved from https://www.klimaatakkoord.nl/documenten/publicaties/2019/06/28/klimaatakkoord

Olah, C. (2015). *Understanding lstm networks.* Retrieved from http://colah.github.io/posts/2015-08-Understanding-LSTMs/

O'Brien, W., & Gunay, H. B. (2015). Mitigating office performance uncertainty of occupant use of window blinds and lighting using robust design. In *Building simulation* (Vol. 8, pp. 621–636).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rafiq, M., Bugmann, G., & Easterbrook, D. (2001). Neural network design for engineering applications. *Computers & Structures*, *79*(17), 1541–1552.

Rijksdienst voor Ondernemend Nederland. (2019). *Energielabels van woningen, 2007*

- *2018*. Retrieved from https://www.clo.nl/indicatoren/nl0556-energielabels-woningen

Rijksdienst voor Ondernemend Nederland. (2020). *Energielabels*. Retrieved from https://energiecijfers.databank.nl/dashboard/dashboard/energielabels/

Rijksoverheid. (2019). *Klimaatakkoord maakt halvering co2-uitstoot haalbaar en betaalbaar*. Retrieved from https://www.rijksoverheid.nl/actueel/nieuws/2019/06/28/klimaatakkoord-maakt-halvering-co2-uitstoot-haalbaar-en-betaalbaar

Rijksoverheid. (2020a). *Beleid in de gebouwde omgeving*. Retrieved from https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/beleid-gebouwde-omgeving

Rijksoverheid. (2020b). *Energielabel woning*. Retrieved from https://www.rijksoverheid.nl/onderwerpen/energielabel-woningen-en-gebouwen/energielabel-woning

Rijksoverheid. (2020c). *Energieprestatie indicatoren - beng*. Retrieved from https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-en-regels/nieuwbouw/energieprestatie-beng/indicatoren

Ripple, W. J., Wolf, C., Newsome, T. M., Barnard, P., & Moomaw, W. R. (2019). World scientists' warning of a climate emergency. *BioScience*.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, *8*(283-297), 37.

Segers, R., Van den Oever, R., Niessink, R., & Menkveld, M. (2019, May). *Warmtemonitor 2017*.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, *404*, 132306.

Smith, J., & Wallis, K. F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, *71*(3), 331–355.

Spinoni, J., Vogt, J. V., Barbosa, P., Dosio, A., McCormick, N., Bigano, A., & Füssel, H.-M. (2018). Changes of heating and cooling degree-days in europe from 1981 to 2100. *International Journal of Climatology*, *38*, e191–e208.

Spoladore, A., Borelli, D., Devia, F., Mora, F., & Schenone, C. (2016). Model for forecasting residential heat demand based on natural gas consumption and energy performance indicators. *Applied Energy*, *182*, 488–499.

Talebi, B., Haghighat, F., & Mirzaei, P. A. (2017). Simplified model to predict the thermal demand profile of districts. *Energy and Buildings*, *145*, 213–225.

Timmermann, A. (2006). Forecast combinations. *Handbook of economic forecasting*, *1*, 135–196.

Tso, G. K., & Yau, K. K. (2007). Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks. *Energy*, *32*(9), 1761–1768.

van Tricht, P. (2020). personal communication.

Verbai, Z., Lakatos, Á., & Kalmár, F. (2014). Prediction of energy demand for heating of residential buildings using variable degree day. *Energy*, *76*, 780–787.

Wang, J.-L., Xue, L., Zhu, L., Chong, Y. S., et al. (2010). Estimation for a partial-linear single-index model. *The Annals of statistics*, *38*(1), 246–274.

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., ... Zhao, X. (2018). A review of data-driven approaches for prediction and classification of building energy consumption. *Renewable and Sustainable Energy Reviews*, *82*, 1027–1047.

Wilkinson, P., Smith, K. R., Beevers, S., Tonne, C., & Oreszczyn, T. (2007). Energy, energy efficiency, and the built environment. *The lancet*, *370*(9593), 1175–1187.

Zhao, H. X., & Magoulès, F. (2010). Parallel support vector machines applied to the prediction of multiple buildings energy consumption. *Journal of Algorithms & Computational Technology*, *4*(2), 231–249.

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

# Appendices

# A Data

All appendices related to the data section can be found here.

## A.1 Variable description

In this appendix an overview of all variables in the space heating demand data set is given.

| Variable | Description | Scale |
|---|---|---|
| PSEUDO_ID | Building ID | Nominal |
| Demand 2007 | Heat consumption per household in 2007 | Ratio |
| Demand 2008 | Heat consumption per household in 2008 | Ratio |
| Demand 2009 | Heat consumption per household in 2009 | Ratio |
| Demand 2010 | Heat consumption per household in 2010 | Ratio |
| Demand 2011 | Heat consumption per household in 2011 | Ratio |
| Demand 2012 | Heat consumption per household in 2012 | Ratio |
| Demand 2013 | Heat consumption per household in 2013 | Ratio |
| Demand 2014 | Heat consumption per household in 2014 | Ratio |
| Demand 2015 | Heat consumption per household in 2015 | Ratio |
| Demand 2016 | Heat consumption per household in 2016 | Ratio |
| Demand 2017 | Heat consumption per household in 2017 | Ratio |
| Demand 2018 | Heat consumption per household in 2018 | Ratio |
| City | Town name | Nominal |
| Zipcode | Zip code | Nominal |
| Neighbourhood | Neighbourhood | Nominal |
| Municipality | Municipality | Nominal |
| Building year | Building year | Ratio |
| Energy label | Energy label (in 2018) | Ordinal |
| House type | Building type | Nominal |
| House type II | Building type (other categorisation) | Nominal |
| Height | Height class of a building (low or high) | Nominal |
| Building surface | Surface of building | Ratio |
| House surface | Floor area of the house | Ratio |

## A.2 Data exploration

For heat demand data, two data sets are available: one containing demand for space heating and one containing demand for hot water. To confirm what literature has found, a short investigation will be conducted to verify the lack of a correlation between hot water demand and temperature. In figure 22 you can see the average demand for hot tap water per year set against the number of degree days in the year. The data points also have the corresponding

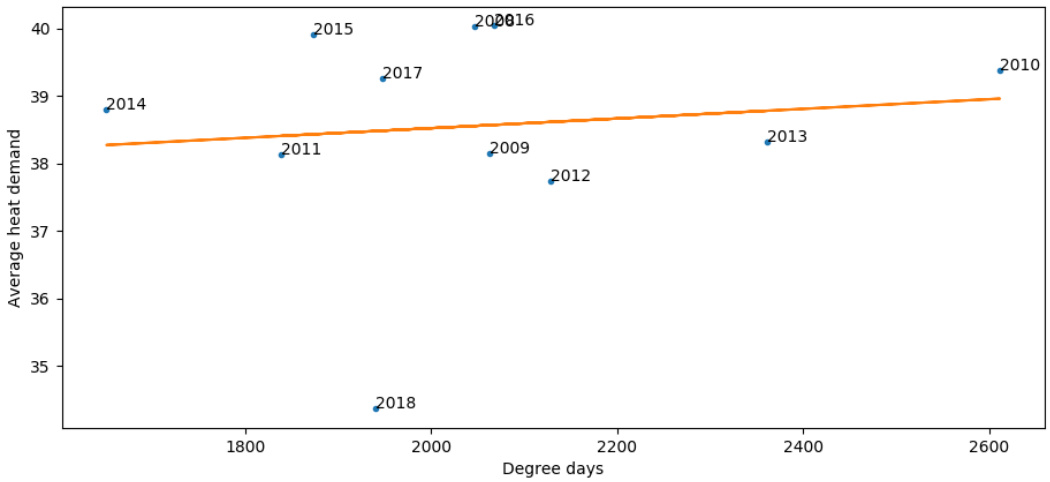


Figure 22: Degree days versus hot tap water demand.

year annotated. The linear regression, which is included in the graph, pointed out that the intercept was 37.1 and significant up to the level of 0.001. The degree days, on the other hand, had an insignificant ($p = 0.736$) coefficient of 0.0007. From this, it is concluded that the number of degree days does not influence hot tap water demand. As a consequence, hot tap water demand is excluded from the analyses and the focus is solely on space heating demand.

In figure 23 you can find the total daily production for the heat network in The Hague set out against the maximum temperature, for days in 2012 (red) and 2016 (blue). This graph makes clear whatever the temperature is, there will always be a certain base demand for heat, the hot tap water demand, that is independent of temperature. When the temperature decreases, the heat demand seems to increase linearly.

These results are supported when regressing the total heat production on temperature. To investigate what threshold temperature for degrees days would perform well, multiple temperatures as threshold are tested. Regressions have been executed for 3 district heating networks in 2012 and for 10 district heating networks in 2016, using only the degree days as regressor. Table 6 shows the averaged $R^2$ values over these regressions for different degree day thresholds. The average and maximum temperature are able to explain more than 80% of the variability. It can be concluded that both the average and maximum temperature thresholds within the range as shown in table 6 are suitable thresholds. Minimum temperatures as a threshold perform clearly worse.

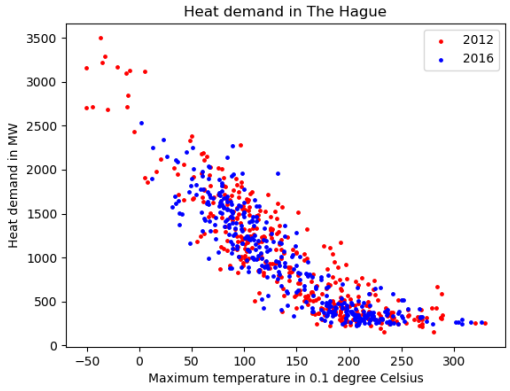


Figure 23: Daily heat demand in The Hague vs. maximum temperature

Table 6: Averaged $R^2$-values for different degree day thresholds.

| Average | | Minimum | | Maximum | |
|---|---|---|---|---|---|
| $Thr$ | $R^2$ | $Thr$ | $R^2$ | $Thr$ | $R^2$ |
| 140 | 0.81 | 90 | 0.64 | 170 | 0.83 |
| 150 | 0.82 | 100 | 0.66 | 180 | 0.84 |
| 160 | 0.83 | 110 | 0.68 | 190 | 0.84 |
| 170 | 0.82 | 120 | 0.69 | 200 | 0.84 |
| 180 | 0.82 | 130 | 0.69 | 210 | 0.84 |
| 190 | 0.81 | 140 | 0.69 | 220 | 0.83 |
| 200 | 0.81 | 150 | 0.69 | 230 | 0.82 |

## A.3 Descriptive statistics

Table 7: Basic descriptive statistics for heat consumption per household in GJ from 2007 to 2018 in the data set before data preparation. In addition the average value per year for data set 1 and 2 are given.

| | Mean | Median | Max | Sd | Mean D1 | Mean D2 |
|---|---|---|---|---|---|---|
| 2007 | 39.4 | 15.0 | 103,023 | 614.4 | 21.39 | 22.23 |
| 2008 | 65.7 | 27.7 | 105,197 | 969.0 | 32.17 | 30.84 |
| 2009 | 44.0 | 27.6 | 18,687 | 239.0 | 31.70 | 30.54 |
| 2010 | 50.0 | 32.6 | 25,250 | 270.5 | 37.19 | 36.23 |
| 2011 | 38,7 | 24.8 | 46,811 | 275.9 | 27.75 | 27.34 |
| 2012 | 42.0 | 26.7 | 100,274 | 466.1 | 29.68 | 29.28 |
| 2013 | 42.8 | 28.4 | 15,785 | 249.1 | 31.66 | 31.31 |
| 2014 | 32.0 | 21.2 | 12,649 | 183.4 | 23.92 | 23.47 |
| 2015 | 35.6 | 23.4 | 16,490 | 205.8 | 26.45 | 26.17 |
| 2016 | 38.0 | 25.3 | 16,100 | 215.4 | 28.49 | 28.07 |
| 2017 | 36.9 | 24.6 | 20,185 | 216.6 | 27.69 | 27.12 |
| 2018 | 38.1 | 24.5 | 113,614 | 458.9 | 27.33 | 26.74 |

## A.4 Missing data

For missing data it is important to make a distinction between different types of missing data. There are three types of missing data (Little & Rubin, 2002):

1. Missing Completely At Random (MCAR): $P(\mathbf{X}_{miss}|\mathbf{X}) = P(\mathbf{X}_{miss})$
2. Missing At Random (MAR): $P(\mathbf{X}_{miss}|\mathbf{X}) = P(\mathbf{X}_{miss}|\mathbf{X}_{obs})$

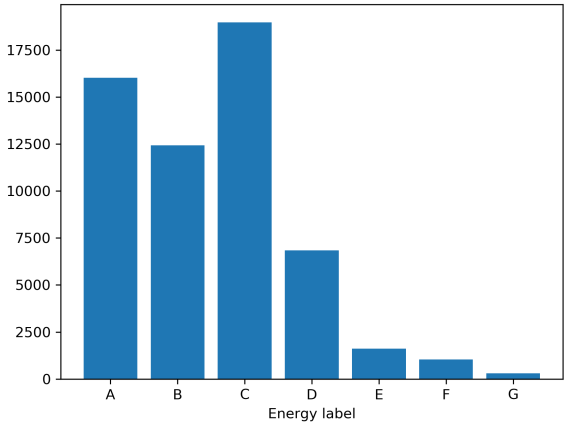


Figure 24: Energy labels in 2018 of houses in the data set.

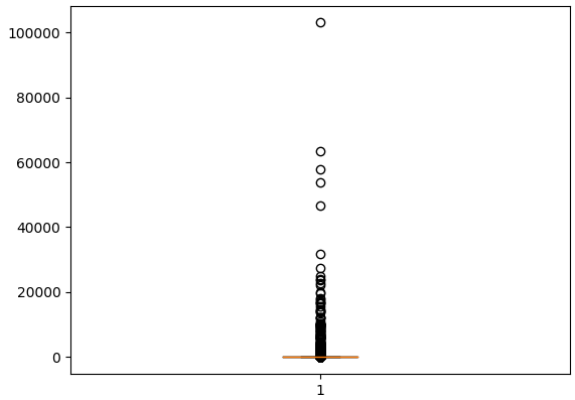


Figure 25: A boxplot of demand values in 2007 (before data preparation).

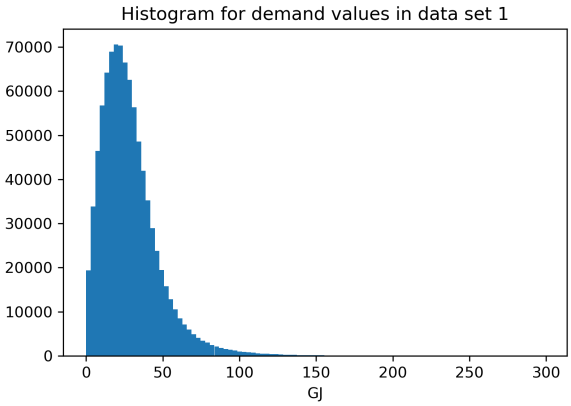


Figure 26: Histogram of all demand values for data set 1.

3. Missing Not At Random (MNAR): $P(\mathbf{X}_{miss}|\mathbf{X}) = P(\mathbf{X}_{miss}|\mathbf{X}_{obs}, \mathbf{X}_{miss})$

In words, the equations above mean the following. MCAR means that the missing data does not depend on the value of either the missing data point itself, or on other values of the same observation. To give an example, if one deals with data about gender and wage, the probability that the wage data is missing, does not depend on the value of wage and does not depend on the value of gender. Hence, it is completely random why a certain value is missing. MAR means that the missing of the data point can be explained with the other values of the same observation. To give an example: if men are less likely to fill in their wage in a survey, the probability that the wage data is missing can be predicted using the observed value of gender. The last option, MNAR, occurs when the probability that a value is missing, depends on the value itself. To use the example again, this means that the probability that wage data is missing, depends on the value of the wage itself.

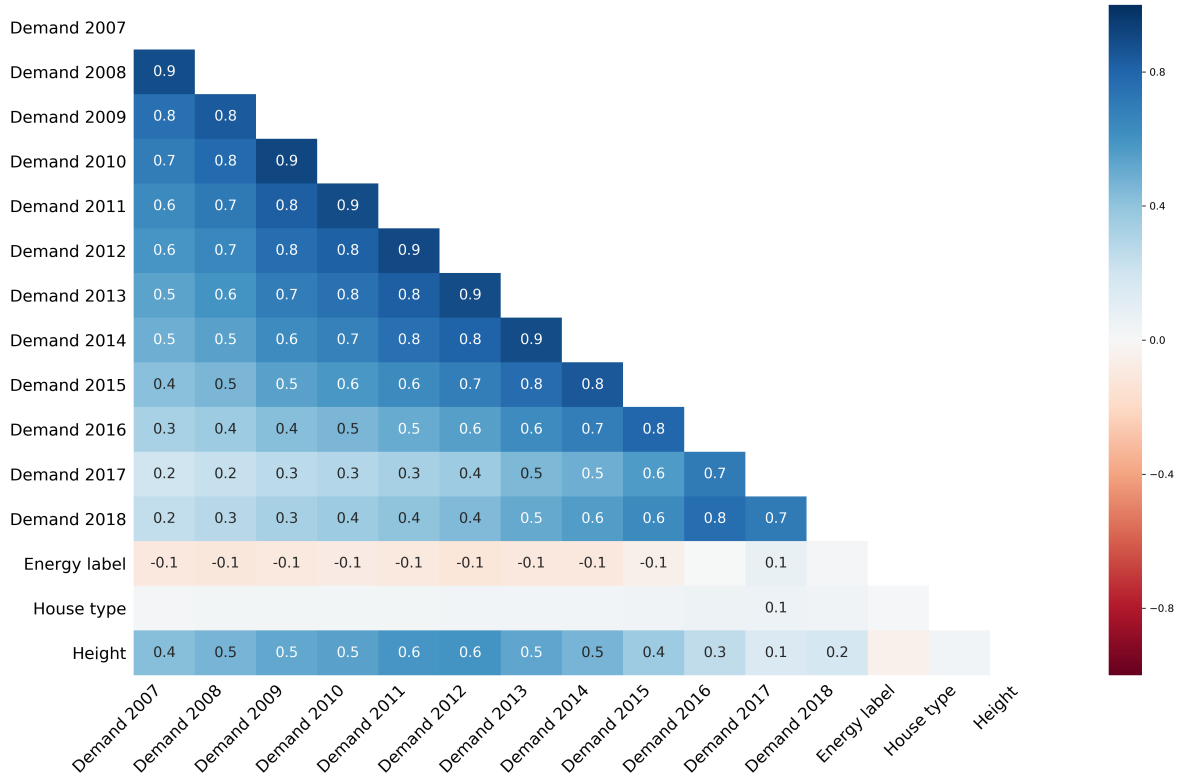


Figure 27: Correlations between missing datapoints per variable.

District heating networks are a collective service for a certain district. It is not possible to connect a different heating district individually, or to switch from heat provider. People who provide in their own heat demand (using heat pumps, solar panels etc.) are very scarce. Hence, usually most houses keep being a client from the moment they are connected to the network. Figure 1 confirms this; the number of clients grows steadily (or more specifically: the number of missing data in space heating demand decreases). The most common reason for missing data is hence that a house was unmanned for a year (apart from data missing from before a house was connected).

It is thus unlikely that the missing of these values depend on building characteristics (this would be MAR). The heat portfolio of Eneco as a whole is influenced by this, because certain types of houses and certain areas are more suitable for district heating networks. However, all observations in the data set are or have been connected to a district heating network. The missing of one or more space heating demand data points thus does not seem to be MAR.

It also does not seem likely that the missing of the space heating demand data depends

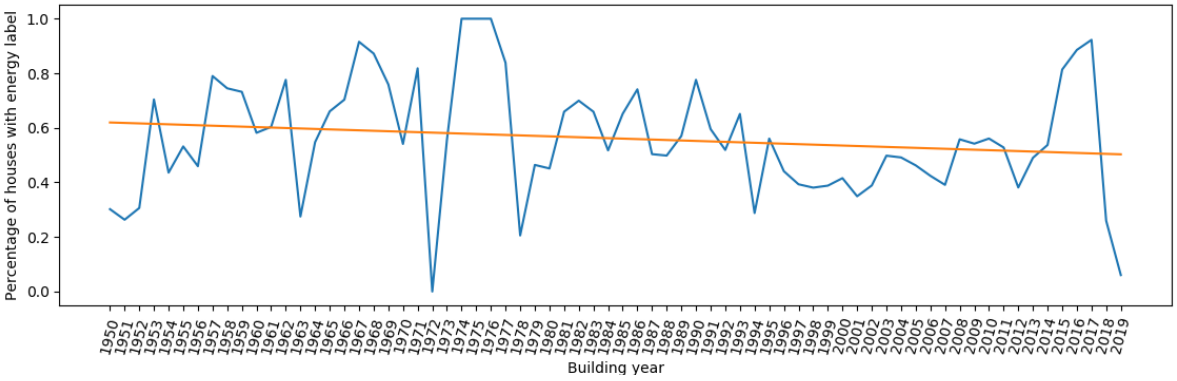


Figure 28: Percentage of assigned energy labels per building year.

on the value itself (MNAR). This would mean that data would be missing from a household if they had an extremely high or low demand for space heating in a certain year. This seems unlikely because the data set is obtained by billing data, and this would not be missing for such a reason. From the paragraphs above, it can be concluded that it is likely that the missing space heating demand data is MCAR.

A serious amount of energy label data is missing as well. Houses get an energy label when they are built, sold or rented (European Parliament and Council, 2002). This would mean that newly built houses are more likely to have an energy label, and thus that the missing data is MAR. However, when taking a look at the building year and the percentage of houses with a label, there is no clear trend (see figure 28). There is no specific reason why households would demand an energy label in another case than one of three above. Hence, it is not expected that, e.g. households with a good energy label are more likely than households with a bad energy label, or vice versa. This would mean that the missing data is MNAR. From the above, it can be concluded that the missing data is MCAR.

## A.5 Scenarios

In this section a short description of the Eneco scenarios is given. The energy label scenarios match these scenarios. Thus, per scenario a short description of the energy label scenarios is presented as well.

### Existing policy

"European and Dutch government policies for meeting 2030 decarbonisation targets are implemented and met" (van Tricht, 2020). Goals for improving the energy efficiency of houses are presented in the so-called Climate Agreement (Nijpels, 2019). The goals and plans of this

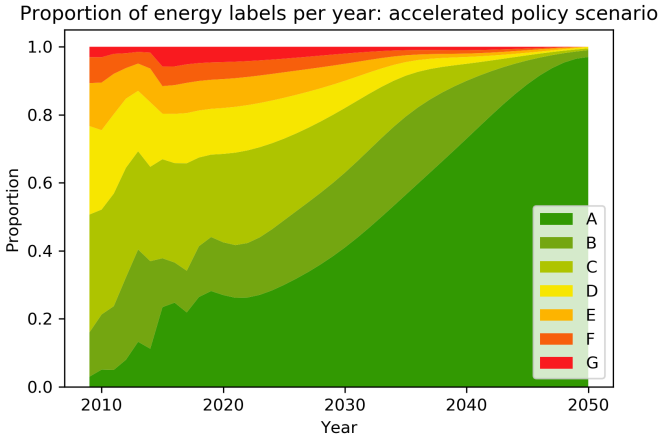


Figure 29: The proportion of energy labels in the accelerated policy scenario.

agreement are taken over by the Dutch government (Rijksoverheid, 2019). In the Climate Agreement, it is stated that in 2021 50.000 houses per year should increase their energy efficiency, and that this should be 200.000 per year in 2030. This forms the basis for the existing policy energy label scenario. In 2021 50.000 houses improve their energy label. This increases linearly until it is 200.000 houses per year in 2030. From 2030 until 2050, 200.000 houses per year improve their energy label. Figure 10 in section 3.5 shows the future distribution of energy labels for the existing policy scenario.

### Accelerated policy

The accelerated policy scenario "envisages the realization of more ambitious climate targets with increased role for offshore wind and a large-scale roll-out of electrolysers" (van Tricht, 2020). The European Union has set a directive for the Netherlands that 300.000 houses or other buildings should improve two label steps every year (Rijksoverheid, 2020a). It is assumed in this ambitious scenarios, that 5 out of 6 of these buildings are houses. Even then, this is a very ambitious goal, and it is not even possible to keep improving energy labels of this many houses up to 2050 (without creating energy labels better than A). As a consequence, it is assumed that from 2030 to 2050, 250.000 houses improve their energy label (and thus not per se with two steps). From 2021 to 2030, there is again a linear increase from 50.000 households that improve their energy label per year in 2021), to 250.000 households per year (in 2030). In the last five years, the number of houses that improve its energy efficiency is below 250.000, because the last houses are known to be very difficult to get to label A (think of monuments, old villas etc.). Figure 29 shows the future distribution of energy labels for the accelerated policy scenario.

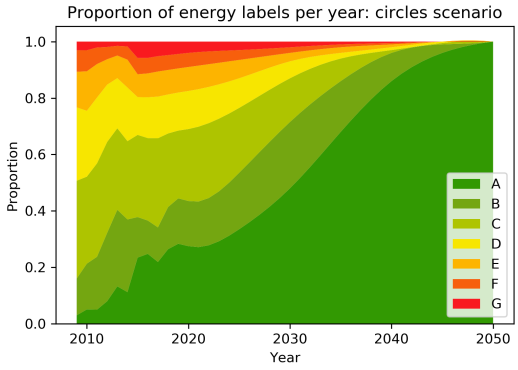


Figure 30: The proportion of energy labels in the circles scenario.

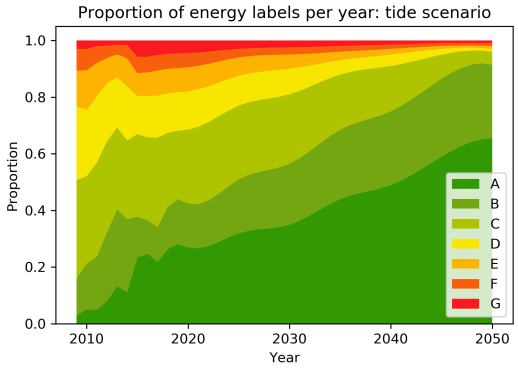


Figure 31: The proportion of energy labels in the tides scenario.

### Circles

"Energy system develops in a decentral way. The energy transition takes place much more rapidly than even the most enthusiastic climate campaigner could have dreamt of" (van Tricht, 2020). This scenario is the most optimistic, and thus an even quicker transition to an energy efficient housing stock is accomplished than in the accelerated policy scenario. The number of houses that improves its label rises more quickly (to 300.000 per year in 2030). From 2035 the number of houses per year that improves its energy label decreases, such that in 2050 the full housing stock has label A. Figure 30 shows the future distribution of energy labels for the circles scenario.

### Tides

"In this highly capitalist world, income is distributed increasingly unequally and the switch to a cleaner energy system is delayed. This results in a future full of volatility and cyclicality in which periods of economic boom and bust alternate" (van Tricht, 2020). As a result, the number of houses that improves its energy label jumps up and down, between 80.000 and 175.000 per year. The idea is that because of the lack of economic stability, it is impossible to reach a peak higher than 175.000. Figure 31 shows the future distribution of energy labels for the tides scenario.

# B Method optimisation

In this section, the optimisation of the three methods is described.

## B.1 Linear panel data models

This section presents the results for the linear panel data models, with a focus on the Random Effects (RE) model. First, this RE model is compared shortly with other linear panel data models: pooled OLS, the fixed effects (FE) model and the between model. Secondly, the predictors to be used in the RE model are determined using forward selection. In this second section, the effects of what appeared to be the most relevant parameters on the space heating demand are presented as well. In the third section, the most promising models for both data sets will be compared.

### B.1.1 Comparison of linear panel data models

The models discussed in this section were compared using both data sets, using the most important regressors where possible: degree days, energy label, building year and surface.

The simplest linear panel data model is pooled OLS (see equation 2). The pooled OLS model does not perform badly, but the RE model performs better. This is always the case, as the RE model is an extension of the pooled OLS model allowing for flexibility per household in the intercept. However, the unadjusted $R^2$ is a lot higher for the RE model for both data sets (see table 8). In addition, the RE model is more efficient in general than the pooled OLS model (exactly because of the household-specific intercept).

The model used for between estimation is shown in equation 4 in section 4.2.1. This is a regression in which the predictors and the demand are averaged over the years for a household. The between estimator exploits thus only differences *between* households. As a result the predictors year and degree days cannot be used, because these are the same for all house-

Table 8: Unadjusted $R^2$-values for Pooled OLS and the RE model using the data set 1 and 2. In addition the residual skewness and kurtosis are given for the RE model.

| | Data set 1 | Data set 2 |
|---|---|---|
| $R^2$ Pooled OLS | 0.21 | 0.33 |
| $R^2$ RE | 0.59 | 0.63 |
| Residual skewness RE | 3.85 | 4.12 |
| Resdiual kurtosis RE | 37.45 | 42.08 |

holds. The between estimator does not perform badly in estimating the average demand per household, having an $R^2$ of 0.488 using data set 2. However, this value cannot be compared to the $R^2$ of the other models, because the averaged demand per household is estimated and not all demand values per household per year. As the RE model can exploit variation in both time and households, this model is preferred above the between model.

The fixed effect estimator has the opposite limitation of the between estimator: it can not include parameters that differ between households but that are constant for households. This, however, is the case for many of the predictors (such as energy label, type of building, and floor surface). It is thus expected to perform badly, and it does indeed perform a lot worse than the RE model. For data set 2, the FE model reaches and $R^2$-value of 0.0632, versus 0.1237 for the RE model. Note that the $R^2$-values here are adjusted $R^2$ values, which measure the explained variation after using the household-specific intercept. As a consequence these values are lower than the unadjusted $R^2$ mentioned earlier. The fixed effects model assigns higher parameter values for the predictor degree days. The t-values for the parameters in the fixed effects models are much lower however (31.6 for FE vs 210.8 for the RE-model using data set 1), so the parameter estimates in the RE model are preferred.

To conclude, the RE model seems to be the best linear panel data model for the data. The RE model outperformed the panel OLS on the unadjusted $R^2$. In addition, the RE model outperformed the between estimator, which was to be expected because the between estimator does not take into account all information. The RE also performed better than the FE model, as it reaches a higher adjusted $R^2$-value. Having decided this, an inspection of the residuals is executed.

The hypothesis that the residuals are normally distributed can be rejected for both data sets, using D'Agostino and Pearson's test for normality. The residual distributions have a skewness around 4 and a kurtosis around 40 (both given in table 8), which is far from what these statistics are in a normal distribution (for which these statistics are 0 and 3 respectively). Hence, a White test for heteroskedasticity is executed. For both data sets, the null hypothesis of normally distributed residuals is rejected for $\alpha = 0.001$. As a result, the RE model is estimated using the heteroskedasticity consistent White estimator.

### B.1.2 Selecting predictors

The forward selection gave almost identical results for both data sets. Table 9 shows the predictors that improved both the AIC and the BIC. It is remarkable to see that the house

Table 9: Predictors that decreased AIC and BIC compared with the base model, ranked by the size of the decrease. In this table the dot between predictors means the interaction effect of these predictors.

| | Data set 1 | Data set 2 |
|---|---|---|
| Surface·Degree days | 1 | 1 |
| Energy label ·Year | 2 | 2 |
| Year·Degree days | 3 | 3 |
| Degree days$^2$ | 4 | 8 |
| Year$^2$ | 5 | 5 |
| Year | 6 | 4 |
| Building year·Year | 7 | 6 |
| Energy label·Degree days | 8 | 9 |
| Building year·Degree days | 9 | |
| Surface·Year | 10 | 7 |

Table 10: The set of predictors included in the different models. In this table the dot between predictors means the interaction effect of these predictors.

| | Data set 1 | Data set 2 |
|---|---|---|
| Full | Degree days, Surface, Building year, Energy label, Surface·Degree days, Building year·Year, Energy label·Year, Year·Degree days, Degree days$^2$, Energy label·Degree days Surface·Year | Degree days, Surface, Building year, Energy label, Surface·Degree days, Building year·Year, Energy label·Year |
| Base | Degree Days, Surface, Building year, Energy label, Surface·Degree days, Building year·Year, House type | Degree Days, Surface, Building year, Surface·Degree days, Building year·Year |

type is not included, while the expectation would be that this has a strong effect on the space heating demand. A possible reason for this could be the bad quality of these data.

Adding all these predictors together improves the quality (measured by AIC and BIC) of the model. The quality even further improves when deleting some predictors. Because of the interaction terms, the models performing best on AIC and BIC can have counter intuitive parameters. Hence, two models are taken into account when comparing the data sets: one that performs best on AIC and BIC (the full model), and one that is easier for interpretation (the base model). Table 10 in appendix B shows the included predictors in these four models.

As the full models include a considerable number of interaction terms, it is useful to have

Table 11: The model quality indicators for the main RE models and the parameter values per predictors. As the parameter values depend on the values of other predictors, the average value for all these predictors are taken..

| | | $\alpha$ | Degree days | Surface | Building year | Energy label | Year | MSE | MSPE | MAPE | MAPPE |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D1 | Base | 112.9 | 0.015 | 0.23 | -0.06 | 1.24 | -0.06 | 177.5 | 335.1 | 1.72 | 1.50 |
| | Full | 53.5 | 0.046 | 0.23 | -0.06 | 1.24 | -0.38 | 174.0 | 330.3 | 1.68 | 1.34 |
| D2 | Base | 183.3 | 0.014 | 0.25 | -0.10 | 0.00 | -0.08 | 183.9 | 337.4 | 2.31 | 1.48 |
| | Full | 123.1 | 0.014 | 0.26 | -0.07 | 1.69 | -0.11 | 183.5 | 337.1 | 2.32 | 1.52 |

a closer look at the effects of different predictors on the space heating demand. Table 11 presents the relative effects for different models. The average values of the other predictors from the data are used (this has to be done because of the interaction terms). In general, the effects are similar. However, the full model has a much higher parameter for degree days. It looks like this is compensated with a much lower constant $\alpha$, and a more negative parameter for year. Apart from the building typology, or house type, all models contain the predictors that were expected (i.e. they have a significant effect) and the effects of the predictors have the expected sign. Hence, for both data sets, a simpler and more easy to understand model and a more extensive model are compared to finally select one model for this method.

The different house types had expected coefficients as well. Detached houses are expected to have a space heating demand of 7.4 GJ more than the reference group and rest category 'Other house'. Corner houses and semi-detached houses have a coefficient of little below 1 (and thus are expected to have 1 GJ more demand than the rest category). Terraced houses have a coefficient of minus 2.0, houses in flats have a coefficient of 2.6 and staircase entrance flats ('portiekwoningen') have an expected demand of 3.7 GJ lower than the rest category. These coefficients are in the expected order, apart from the fact that flat was expected to have a lower demand than staircase entrance flats.

### B.1.3 Comparing the RE model for the two data sets

The four models are compared using the training and validation set. For the four models, the main performance indicators are presented in table 11. Although data set 2 has better $R^2$-values (see table 8), data set 1 performs better in prediction. The full models perform better than the basic models. That the models perform a lot worse on the MSPE than the MSE is caused by the household specific intercept, which cannot be included in prediction. It is remarkable though, that the models perform better on the MAPPE than on the MAPE.

From table 11 it could be concluded that the full model for data set 1 performs best.

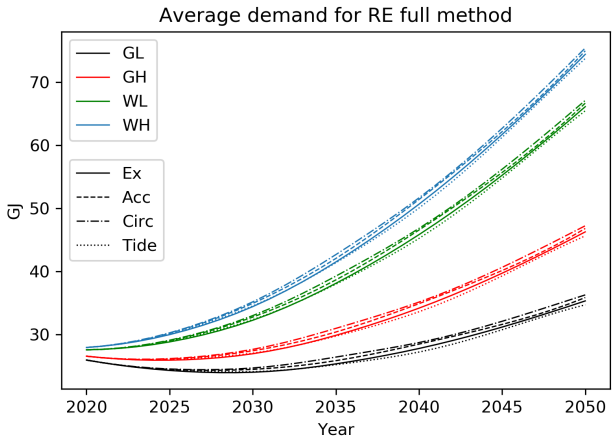


Figure 32: The development of the average heat demand in different scenarios for the full RE model. The space heating demand decreases in the first few years for most scenarios, but quickly starts rising.

However, when forecasting, the model does not perform well. As already found in table 11, the full model for data set 1 had considerably different relative effects than the other models. The relative effect for degree days becomes negative when the value for year increases. Taking the averages for other predictors, the relative effect of degree day becomes negative from the year 2026 on. This seems incorrect, as all models found a positive parameter for degree days. However, such an incorrect change of the sign of parameter can happen, because the model is not trained on values of year higher than 2026. The result is shown in figure 32: the heat demand increases, while the number of degree days decreases. This is not realistic, and as a consequence, the second best model is used, which is the base model for data set 1.

In addition, the exclusion of house type is double-checked. After it was excluded when looking at the AIC and BIC, it was added to the base model for data set 1 to investigate the prediction performance. The prediction performance (in terms of MSPE) improved including the house type. As the main goal of the research is forecasting, the final RE model is the base model for data set 1 including house type.

## B.2 Functional Coefficient Model

This section presents the results of the Functional Coefficient Model (FCM). First, the results using degree days and surface as varying variable are discussed. The main predictors from the RE model, degree days, floor surface, energy label and building year are included in these models, and year as well. Secondly, an intercept and the house types (a categorical variable) are added. Thirdly, the results of the most promising model are compared using data set 1

and 2.

### B.2.1 Experimenting with the varying variable

Degree days and surface are used as varying variable (separately). When using degree days as varying variable, the model behaves as expected. Stronger effects are found for higher values of degree days, i.e. higher parameter values in absolute terms. This is visible in figure 33, that shows that the coefficients of surface, energy label and year increase strongly in absolute terms. The coefficient for degree days itself increases only slightly for most models and sometimes even decreases slightly. This indicates a linear relationship between degree days and space heating demand. The figure shows that the coefficient for building year stays relatively constant too. The straight appearance of the lines makes clear that a sufficient amount of estimation points is used (the step size was 200 degree days).

When using surface as varying variable, the coefficients have less smooth functions. For high values of surface, the parameter values jump up and down strongly, as becomes clear from figure 34. This behaviour might be caused by the fact that there are very few houses in the (training) set with high values for surface (as is shown in the histogram in figure 35). The few high values have a large impact on the estimated parameters, because the kernel function in equation 10 strongly reduces the impact of data points far from $u_0$, the point at which the parameter is estimated. If these few high values for surface have extreme or unusual values for the endogenous variable demand, this can cause strange behaviour. Next to this strange behaviour, the performance of models with floor surface as varying variable in terms of MS(P)E, MAP(P)E and $R^2$ is worse than the performance of models with degree days as varying variable. This is probably caused by this strange behaviour of the coefficients. Because of this bad performance, degree days is used as varying variable in the following sections.

### B.2.2 Intercept and house type

Adding an intercept depending on the varying variable only works for models where the varying variable is not included in the set of predictors. This is caused by the fact that an intercept with a value depending on the varying variable can behave exactly the same as the varying variable that is included in the set of predictors. The value of the intercept itself is simply 1. Hence the value of its coefficient at $u_0$, $a_{int}(u_0)$, times its own value, 1, is equal to the coefficient $a_{int}(u_0)$. The value of $a_u(u_0) * u$ can then be exactly the same for all values of $u_0$ under the assumption that equation 13 is correct. The model is thus overspecified in case that both the varying variable and an intercept are included in $\mathbf{X}$.

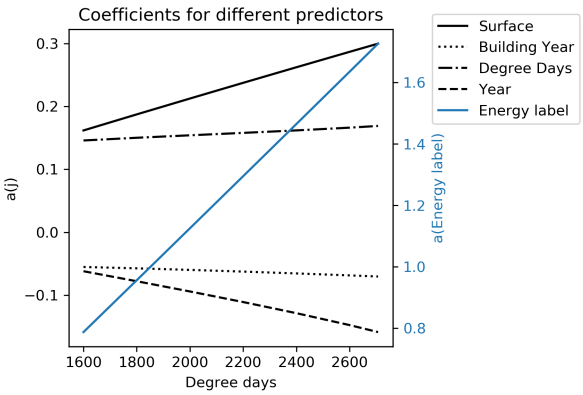


Figure 33: Coefficients for the predictors surface, energy label, building year, degree days and year, with degree days as varying variable for data set 1. Note that the value of energy label (in blue) corresponds to the right y-axis (also in blue).

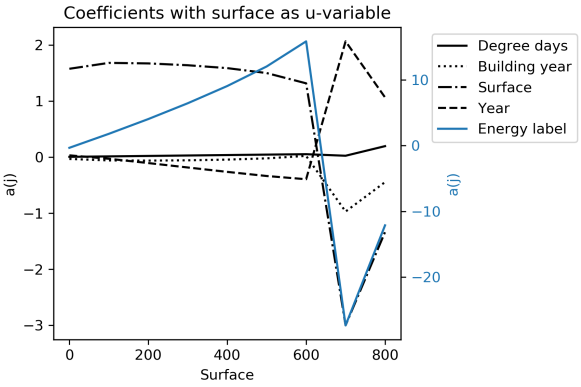


Figure 34: The coefficients show strange behaviour for high values of surface with surface as varying variable. Note that the value of energy label (in blue) corresponds to the right y-axis (also in blue).

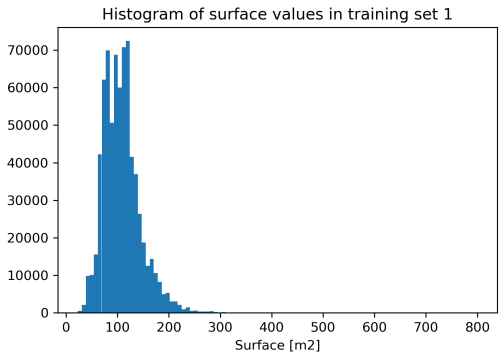


Figure 35: The histogram of surface is very skewed. There are so few high values compared to low values, that they cannot be identified in this graph.

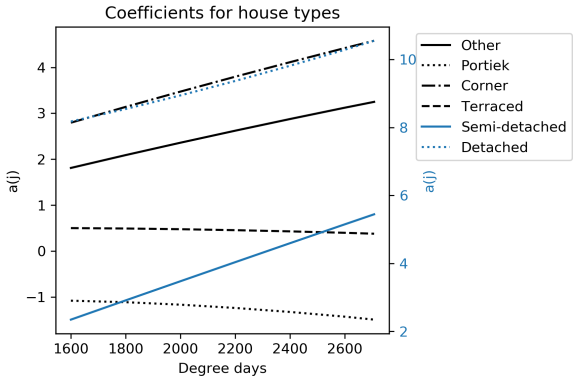


Figure 36: Coefficients for different house types for data set 1 and with degree days as varying variable. The blue lines corresponds to right y-axis.

Table 12: Comparing the inclusion of the varying variable, always degree days in this case, and inclusion of the intercept. Whether the varying variable is included in **X** or the intercept is included is shown in the second column. The models have the same performance on all indicators in the table.

| Variables in **X** | | MSE | MAPE | MSPE | MAPPE |
|---|---|---|---|---|---|
| Building year, surface, | U | 336.3 | 1.46 | 338.1 | 1.42 |
| Energy label | int | 336.3 | 1.46 | 338.1 | 1.42 |
| Idem, plus house type | U | 334.3 | 1.45 | 336.3 | 1.41 |
| | int | 334.3 | 1.45 | 336.3 | 1.41 |

$$a_u(u_0) = \frac{a_{int}(u_0)}{u} \tag{13}$$

As a result, it is not possible to include both an intercept and the varying variable as predictor. Using the varying variable as part of the set of predictors logically results in the same performance as including an intercept on the performance indicators. This is shown in table 12. It makes more sense to use the variable that actually influences the output instead of a non-constant intercept depending on a predictor. As a consequence, degree days is included in the set of predictors in **X**.

Including the house type improves the model performance, although generally differences in performance between the models are small (see table 13). The coefficients per house type in general show the same behaviour as coefficients of other parameters: positive ones tend to increase with the increase of degree days and negative ones tend to decrease with the increase of degree days. In addition, houses surrounded by more open space have higher coefficients, as expected (such as detached and semi-detached houses). Both these effects are visible in figure 36. It is remarkable that a staircase entrance flat ('portiekwoning') has a lower coefficient than 'flat'. Looking only at open space around a house, a flat would be expected to have a lower space heating demand. However, wind could possibly play a role, because houses in flats are higher than staircase entrance flats, and thus catch more wind.

### B.2.3 Parameter estimation

It might seem that the parameter estimates are straight lines (in figures 33 and 36). This, however, is not the case, although for some predictors the lines are almost straight. When the lines are shown separately in a figure, however, it becomes clearer that the lines are indeed not straight (which could be a sign that something went wrong in the estimation). Figure 37 shows this for the predictor terraced house.

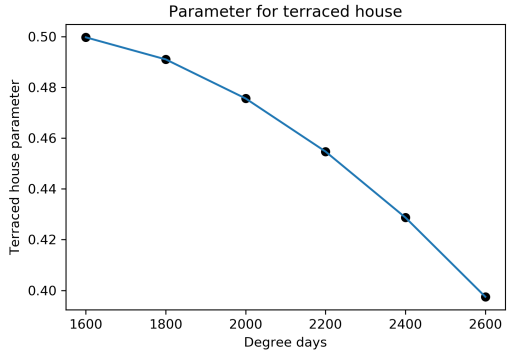


Figure 37: The parameter values for a terraced house, depending on the value of degree days.

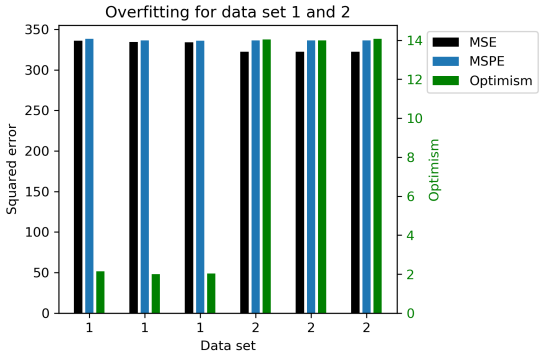


Figure 38: MSE, MSPE and the overfitting for models from data set 1 and from data set 2.

The parameter estimates seem to be reasonably reliable in general. The variation in the parameter estimates using 30 different bootstrap samples are generally small. However, the relative standard deviations (i.e. the standard deviation divided by the mean) differ substantially per predictor. For floor surface, building year and degree days, the relative standard deviation is below 5% for all values of degree days. For energy label this is already slightly higher (between 3 and 11%), and for the house types the insecurity is even bigger. The relative standard deviations are generally below 20%, excluding the one of terraced house. The large differences in the coefficient estimate of terraced house are shown in figure 40. On the other hand, figure 39 shows the reliability of the estimates for floor surface. The figures for degree days and building year look similar. In general, it is also found that the parameter estimates are more variable for extreme degree day values, as can also be seen from figures 39 and 40.

Although the parameter estimates can thus vary sometimes, the predictive quality of the model with different training data is very stable. For the thirty different sets, the MSPE varies only from 332.8 to 333. This indicates that the instability of some coefficients does not affect the model performance.

### B.2.4 Comparison data sets and models

For both data sets, the best models include the degree days, house type, surface, energy label and building year in $\mathbf{X}$, and have degree days as varying variable. The models for data set 1 again perform slightly better when predicting out-of-sample. With the MSPE and MAPPE around 334 and 1.5 respectively, the performance of the functional coefficient model is comparable with that of the RE model. It is remarkable to see that the MSE is lower for data set 2, but that the MSPEs are similar (see figure 38). Naturally, optimism is the result. It

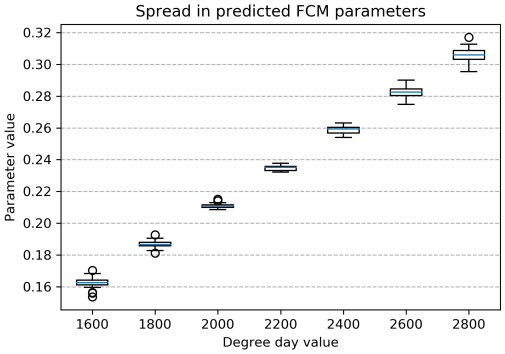


Figure 39: The parameter estimates for floor surface for different degree day values.

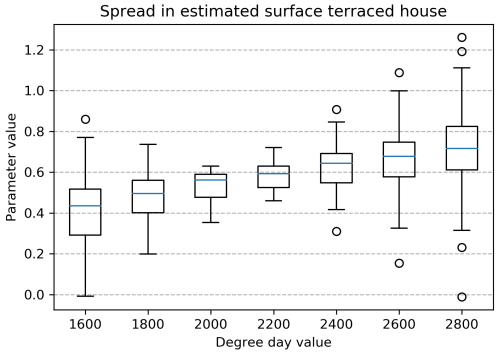


Figure 40: The parameter estimates for terraced house for different degree day values.

Table 13: Prediction performance of best models for data set 1 and 2. All models have degree days as varying variable.

| | MSPE | MAPPE | Variables in **X** |
|---|---|---|---|
| Data set 1 | 332.9 | 1.56 | Surface, Energy label, House type, Building year, Degree days |
| | 334.7 | 1.48 | Idem, plus Year |
| Data set 2 | 333.6 | 1.58 | Surface, Energy label, House type, Building year, Degree days |
| | 336.3 | 1.50 | Idem, plus Year |

thus seems that data set 2 is more prone to overfit. As the model without year for data set 1 has the lowest MSPE, this is the preferred FCM model. Because data set 1 performed better for the linear panel models and for the functional coefficient model, data set 2 is from now excluded.

## B.3 Recurrent Neural Network

All neural networks include the variables that appeared most valuable from the foregoing methods. These are degree days, building year, energy label, floor surface and house type.

### B.3.1 Long short-term memory

The long short-term memomy (LSTM) performed worse than expected. Looking at the loss functions over multiple epochs, it becomes clear that for all hyperparameter settings only the loss function of the training set decreases, while the loss function of the validation set quickly increases. This often happens already in the first few epochs, as is shown in figure 41. The

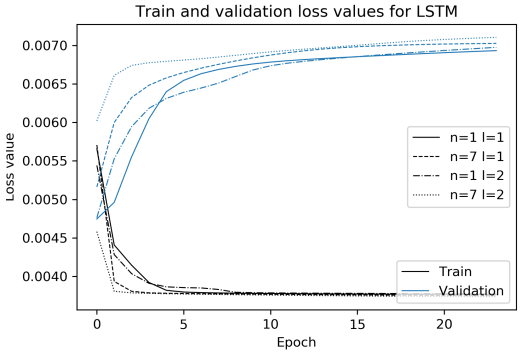


Figure 41: The train and validation losses (for normalised data), averaged over 15 runs for LSTM models with different settings for the number of nodes (n) and layers (l). The batch size was 48 for all models.

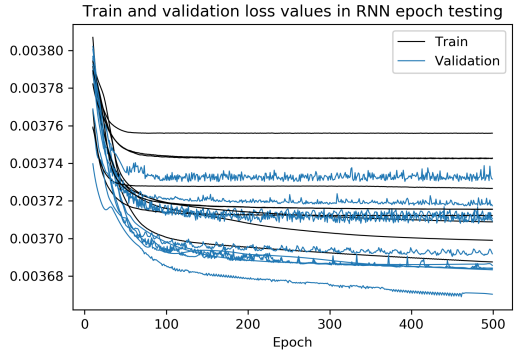


Figure 42: The loss values for the training and validation set decrease sharply in the first epochs, but plateau after about 100 epochs.

MSPEs are much higher than those of the RE and FCM model. Even when experimenting with many different hyperparameter settings, the best result is an MSPE of 585. The RE and FCM models obtained MSPEs of 335 and 332 respectively for predicting the validation set. As the quick increase of the validation loss function made clear, the model performs better with fewer epochs. However, even with only a few epochs, the MSPEs are worse than those for the other methods.

The increase of the validation loss function and decrease of the training loss function are signs of overfitting. In this case, the overfitting happens after very few epochs. Comparing different hyperparameter settings also leads to suspicions of overfitting. The less complex models, with fewer nodes and layers, perform better than more complex models, as the loss values in table 14 show. Figure 41 also shows that the less complex models perform slightly better in the first few epochs. It can be concluded that the LSTM finds complex relations in the data that do not exist.

### B.3.2 Regular Recurrent Neural Network

The regular recurrent neural network (RNN) performs better than the LSTM. Running the model for a large number of epochs shows that the loss functions for the training and validation strongly decrease in the first epochs. The improvements become smaller after 20 epochs. After 100 epochs the loss values plateau in general (see figure 42). The optimal number of epochs is thus 100.

Table 14: The average loss value of the test set after one epoch. For every hyperparameter setting, 15 runs are executed. Experiments are executed with all possible combinations of the batch sizes, number of layers and number of nodes as in the table. Thus, the average value for nodes is the average of 15 runs for 8 (4 batch size options and 2 layer options) hyperparameter settings. LSTM cells were used for these runs. Note that the table shows normalised loss values.

| | Value | Average test loss |
|---|---|---|
| Nodes | 1 | 5.07 E-03 |
| | 3 | 5.49 E-03 |
| | 5 | 5.54 E-03 |
| | 7 | 5.79 E-03 |
| Layers | 1 | 5.42 E-03 |
| | 2 | 5.52 E-03 |
| Batch size | 6 | 6.31 E-03 |
| | 12 | 5.90 E-03 |
| | 24 | 5.64 E-03 |
| | 48 | 5.10 E-03 |
| | 96 | 4.88 E-03 |

However, due to large run times, the following hyperparameter tests are compared among each other with a smaller number of epochs (50). The batch sizes tested are 24, 48, 96 and 192. The batch sizes are tested with one layer and 5 nodes. Although the average MSPE of models with batch size 24 is slightly lower, the model with batch size 96 is considered best. This is a result of the smaller spread in MSPE and the run time, which is more than three times shorter than the run time of the model with batch size 24.

Because of the overfitting of LSTM, relatively simple models are used, with only one or two layers and 3, 5, 7 or 9 nodes per layer. The model with two layers and 7 nodes per layer has the smallest MSE and MSPE, but the model with two layers and 9 nodes per layer obtains similar results. As a consequence, these two models are tested using 100 epochs. All results for the hyperparameter testing of the RNN are shown in table 15.

Table 15: The averaged results over 10 runs for different hyperparameter settings in the RNN experiments. These results were found using regular RNN cells.

| Exp | Epochs | Batch size | Layers | Nodes | MSE | MAPE | MSPE | MAPPE | O | MSPE min | MSPE max | MSPE sd | Runtime per epoch [s] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 50 | 24 | 1 | 5 | 331.3 | 1.73 | 329.1 | 1.89 | -2.14 | 325.5 | 333.4 | 2.79 | 34 |
| 2 | 50 | 48 | 1 | 5 | 333.9 | 1.77 | 331.9 | 1.94 | -2.00 | 330.8 | 332.6 | 0.60 | 18 |
| 2 | 50 | 96 | 1 | 5 | 330.8 | 1.73 | 329.2 | 1.89 | -1.56 | 328.0 | 331.3 | 1.02 | 10 |
| 2 | 50 | 192 | 1 | 5 | 331.6 | 1.71 | 331.2 | 1.87 | -0.44 | 329.0 | 333.7 | 1.60 | 5 |
| 3 | 50 | 96 | 1 | 3 | 331.5 | 1.73 | 329.9 | 1.90 | -1.63 | 328.3 | 332.3 | 1.58 | 10 |
| 3 | 50 | 96 | 1 | 5 | 330.9 | 1.73 | 329.5 | 1.89 | -1.47 | 327.9 | 330.9 | 1.03 | 10 |
| 3 | 50 | 96 | 1 | 7 | 331.1 | 1.73 | 329.5 | 1.90 | -1.56 | 328.2 | 332.0 | 1.33 | 10 |
| 3 | 50 | 96 | 1 | 9 | 331.6 | 1.74 | 330.0 | 1.90 | -1.54 | 328.5 | 333.2 | 1.34 | 10 |
| 3 | 50 | 96 | 2 | 3 | 330.3 | 1.74 | 328.3 | 1.90 | -1.98 | 327.0 | 329.8 | 0.99 | 13 |
| 3 | 50 | 96 | 2 | 5 | 328.2 | 1.74 | 326.1 | 1.90 | -2.16 | 324.0 | 328.0 | 1.27 | 13 |
| 3 | 50 | 96 | 2 | 7 | 327.8 | 1.73 | 325.6 | 1.90 | -2.22 | 324.1 | 328.7 | 1.27 | 13 |
| 3 | 50 | 96 | 2 | 9 | 328.0 | 1.73 | 325.7 | 1.90 | -2.32 | 324.7 | 326.9 | 0.81 | 13 |
| 4 | 100 | 96 | 2 | 7 | 325.9 | 1.74 | 323.8 | 1.90 | -2.12 | 323.2 | 325.6 | 0.70 | 13 |
| 4 | 100 | 96 | 2 | 9 | 326.4 | 1.74 | 324.0 | 1.90 | -2.41 | 323.1 | 326.2 | 1.12 | 13 |

When comparing the final two models over more epochs (100), slightly lower means squared prediction errors are found (around 324 vs around 325.5). The model with 7 nodes has a slightly lower average MSPE and in addition a smaller spread of MSPE. This is thus the RNN model that is considered best based on the hyperparameter testing.

# C Parameter comparison

Table 16: This table shows the parameter values for the RE and functional coefficient model for different splits in the data set. The coefficients of the FCM are taken for the degree days value 2200, which is closest to the average value. As the coefficients are relatively straight lines, taking the average value is the most valid. Set shows the relative size of the training set.

| Method | Set | int | Surface | Energy label | Building year | Degree days |
|---|---|---|---|---|---|---|
| RE | 80 | 122.1 | 0.20 | 1.22 | -0.063 | 0.013 |
| FCM | 80 | | 0.24 | 1.25 | -0.064 | 0.058 |
| RE | 85 | 121.8 | 0.20 | 1.22 | -0.062 | 0.013 |
| FCM | 85 | | 0.24 | 1.24 | -0.064 | 0.058 |
| RE | 90 | 118.9 | 0.20 | 1.26 | -0.061 | 0.013 |
| FCM | 90 | | 0.24 | 1.29 | -0.062 | 0.056 |
| RE | 95 | 118.0 | 0.20 | 1.30 | -0.061 | 0.013 |
| FCM | 95 | | 0.24 | 1.33 | -0.062 | 0.056 |

Table 17: This table shows the parameter values for the RE and functional coefficient model for different splits in the data set. The coefficients of the FCM are taken for the degree days value 2200, which is closest to the average value. As the coefficients are relatively straight lines, taking the average value is the most valid. Set shows the relative size of the training set.

| Method | Set | Other | Porch | Corner | Terraced | Semi-detached | Detached |
|---|---|---|---|---|---|---|---|
| RE | 80 | 2.46 | -1.16 | 3.78 | 0.47 | 3.97 | 9.53 |
| FCM | 80 | 2.51 | -1.18 | 3.87 | 0.48 | 4.09 | 9.59 |
| RE | 85 | 2.68 | -1.09 | 3.62 | 0.45 | 3.91 | 9.00 |
| FCM | 85 | 2.74 | -1.11 | 3.69 | 0.45 | 4.02 | 9.06 |
| RE | 90 | 2.52 | -1.27 | 3.76 | 0.49 | 3.92 | 9.24 |
| FCM | 90 | 2.56 | -1.30 | 3.84 | 0.50 | 4.04 | 9.30 |
| RE | 95 | 2.60 | -1.18 | 3.62 | 0.45 | 3.88 | 9.24 |
| FCM | 95 | 2.66 | -1.20 | 3.69 | 0.45 | 4.01 | 9.30 |