# The prediction of actual energy use during the use phase of Dutch dwellings using building specific parameters.

G.B. Eggens BSc.
S.1251090
Master Environmental and Energy Management
Academic year 2018/2019
University of Twente

In cooperation with: Royal BAM group

Supervisors:
dr. Y. Krozer
dr. M.L. Franco Garcia

Supervisor of BAM:
W. Schakel

## Abstract

The building sector contributes significantly to greenhouse gas emissions. To reduce the energy related $CO_2$ emissions which occur during a building's lifecycle, knowledge is needed on where the emissions occur. This study aims to tackle this knowledge gap by investigating if building specific parameters can be used to predict the actual energy use during the use phase of Dutch dwellings. In this explorative study to predict the actual energy use this study first assesses current methods used in literature and subsequently uses cross-validated stepwise multiple linear regression on the 'Woon Onderzoek Nederland' [1] dataset using only dwellings built after 2007 and using only building specific parameters. The building type, the theoretical total energy use ('Energy Performance Certificate' score combined with the gross floor area), and the number of rooms (in 5 classes) are identified in the multiple linear regression analysis as the key (building specific) parameters in predicting the actual energy use during the use phase of Dutch dwellings. The model created to predict the actual energy use during the use phase of Dutch dwellings shows that 35% of the variance in the data can be explained with these building specific parameters.

# Content

# List of Tables, Figures and Graphs

## List of Acronyms

| | |
|---|---|
| BAM CME | BAM Construction M&E services |
| BIM | Building Information Modelling |
| BZK | Government of the Netherlands (Binnenlandse Zaken en Koninkrijksrelaties) |
| CBS | Statistics Netherlands (Centraal Bureau voor de Statistiek) |
| CRM | Customer Relationship Management |
| CV | Central heating boiler (Cetrale Verwarming) |
| DV | Dependent Variable |
| EIO | Economic Input-Output |
| EPC | Energy Performance Coefficient (Energie Prestatie Coeffieciënt) |
| EPI | Energy Performance Index |
| GFA | Gross Floor Area |
| GHG | Greenhouse Gas |
| IEA | International Energy Agency |
| IV | Independent Variable |
| LCA | Life Cycle Assessment |
| $LCCO_2A$ | The Life-Cycle Carbon Emission Assessment |
| LCEA | Life Cycle Energy Assessment |
| SQ | Sub Question |
| VIF | Variance Inflation Factor |
| WBCSD | World Business Council for Sustainable Development |
| WoON | Woon Onderzoek Nederland |

## Acknowledgement

During the process of working on this research project, I've learned a lot. I could not have done it without the help of others. Teachers, fellow students, friends and family all have helped me a lot during this time.

First of all I would like to thank my first supervisor dr. Yoram Krozer for the interesting meetings we had via Skype, and for the feedback and suggestions provided. Especially the direct comments and suggestions on the work written. Also a special thanks with the help in the beginning for finding the right balance between wishes of BAM and the scientific value of this thesis. I would also like to thank the second reader of my thesis dr. Laura Franco-Garcia for the comments on the work delivered.

During my thesis period I had a combined assignment at BAM. I would like to thank all colleagues at BAM with which I had discussions about several topics, and also for including me in the Sustainability, innovation and strategy team. A special thanks to Wouter, my supervisor from BAM, he helped me with a lot of suggestions for meeting people within BAM and with the readability of this thesis.

Finally I would like to thank my parents for supporting and encouraging me to do the MEEM program. Next to this I want to thank my girlfriend for all the moral support, reading through several versions of my thesis, commenting on the structure and the help with SPSS.

# 1 Introduction

Current environmental problems such as global warming caused by greenhouse gas emissions have led to the increasing attention for reducing the impact of human activities on climate change [2]. The building and construction sector was responsible for 39% of energy related $CO_2$ emissions in 2016, if upstream power generation is included [3]. So, the building sector contributes significantly to global Greenhouse Gas (GHG) emissions. Moreover the building industry is a major consumer of natural sources, in the EU 50% of raw material consumption is accountable to the built environment [4]. The International Energy Agency (IEA) shows that, with existing policies and commitments, the energy demand of the building sector will increase 30% by 2060 if there is no more ambitious effort to decrease carbon use and increase energy efficiency of construction and buildings [3].

To mitigate the effects of the building industry on climate change and to be able to reduce the $CO_2$ emissions which occur during a building's lifecycle, knowledge is needed on where the emissions occur. This knowledge can be gained by calculating and reporting on emissions. It is important to track and report emissions for companies to understand the impact they have on GHG emissions and climate change [5].

There are several ways to calculate and report GHG emissions on the company level. The GHG protocol is the most widely used and accepted framework for emission reporting [5]. The GHG protocol is the result of a 20-year partnership between World Resource Institute (WRI) and the World Business Council for Sustainable Development (WBCSD). The framework can be used to show, measure and manage the GHG emissions associated with sector operations, value chains and mitigation actions [6]. This tool serves as a standard framework for reporting GHG emissions on the company level [7]. The GHG protocol divides a company's emissions in three categories: direct emissions on site (scope 1), emissions from electricity, heating and cooling (scope 2), and the up and downstream activities in the value chain (scope 3) [8]. The emissions in Scope 1 and 2 are relatively easy to collect in comparison to scope 3. Hertwich et al. [9] show that the total $CO_2$ equivalent emissions in scope 3, for buildings and construction, for the whole sector, are twice as high as the direct emissions in scope 1 and 2. The GHG protocol also provides some guidance to companies to calculate their emissions. However, an exact calculation method is not provided. The protocol does require from its practitioners that the calculation, of the $CO_2$ emissions, improves each year and that the calculation is transparent.

Several researchers show that the operational phase, also called the use phase, of a building is the highest energy consumer [10], [11], [12]. Gong et al. [13] show that the use phase is responsible for 80-90% of the energy consumption over the whole life cycle of a building. In literature the use phase of a building is defined in several ways, the use phase is also called the operation phase or the operational phase. In this study the $CO_2$ emissions of the use phase of a building is defined as the emissions related to the energy use during the operation of a building. So, embodied carbon, water use, maintenance, repair, replacement, and refurbishment are excluded in this study.

The $CO_2$ emissions in scope 3 are split in 15 categories are also analyzed by a contractor, Royal BAM. They identified, in an explorative study, the 'use of sold products' and the 'purchased goods and services' category as the biggest $CO_2$ emitters. This study will focus on this 'use of sold products' category. The sold products for a contractor are buildings and civil structures, since the emission calculations for the use phase of civil structures are complex it is not clear how these emissions could be estimated.

To demarcate this study, the focus will be on the emissions which occur due to energy use during the use phase of residential buildings.

There are several existing approaches which can be used to calculate the emissions during the operation of residential buildings. There is not yet one standard, time-efficient approach. This study aims to fill this gap in two-fold: by exploring existing approaches to calculate emissions during the use phase of residential buildings, and by identifying key parameters in emission calculations to predict the actual energy use during the operation phase of residential buildings.

**The main research question:**

*How can the energy related emissions during the use phase of a Dutch dwelling be predicted using building specific data?*

Sub-questions:

1. *What approaches are available for the $CO_2$ emission calculations of the use phase of dwellings?*
2. *What are the key parameters in predicting the actual energy use of a dwelling?*
3. *How are approaches different in terms of data need?*
4. *What is the accuracy of the approaches?*
5. *How can an approach be used for applying it at a contractor?*[1]

The first sub-question (SQ) will be answered in chapter 2, the theoretical framework. Chapter 2 will also be used to partly answer SQ 2, 3 and 4. To answer SQ 2 properly the WoON dataset is used to perform a multiple regression analysis to identify statistically significant predictors of the actual energy use. Chapter 3 will elaborate on the study design, the method of data collection, processing and analysis, to explain the multiple regression analysis which will be used to answer SQ 2, 3 and 4. The results are presented in Chapter 4 and will provide answers to SQ 2, 3 and 4. Chapter 5 will present a case study to show the applicability of the method at a contractor which will answer SQ 5. Chapter 6 discusses the results and limitations of this study. The conclusion is presented in chapter 7 to answer the sub-questions and the main research questions.

[1] This study was performed in cooperation with Royal BAM. Therefore, this question looks at the implementation at a contractor, which wants to improve its emission reporting.

# 2 Theoretical framework

There are several approaches used in literature to calculate the emissions of the use phase of residential buildings. This section elaborates on each of these approaches. Criteria for selecting the approaches were, that the methods take building´s properties into account and translate this to $CO_2$ emissions or energy use.

First existing approaches to predict the actual energy use of dwellings, followed by section on important parameters in emission calculations of the use phase of buildings, identified in other studies. This chapter finishes with methods to identify important predicting parameters of the actual energy use of dwellings.

## 2.1 Existing approaches

This section describes the existing approaches which are used to predict the actual energy use of dwellings.

### 2.1.1 Life Cycle Assessment

Life Cycle Assessment (LCA) is a method to assess all the emissions related to a product or a process, which takes the whole life cycle into account, e.g. environment & resource depletion, from cradle to grave [14]. Four steps are necessary to perform a LCA: the goal and scope definition, the life cycle inventory analysis, the life cycle impact assessment, and the interpretation, which is visualized in Figure 1 [15].

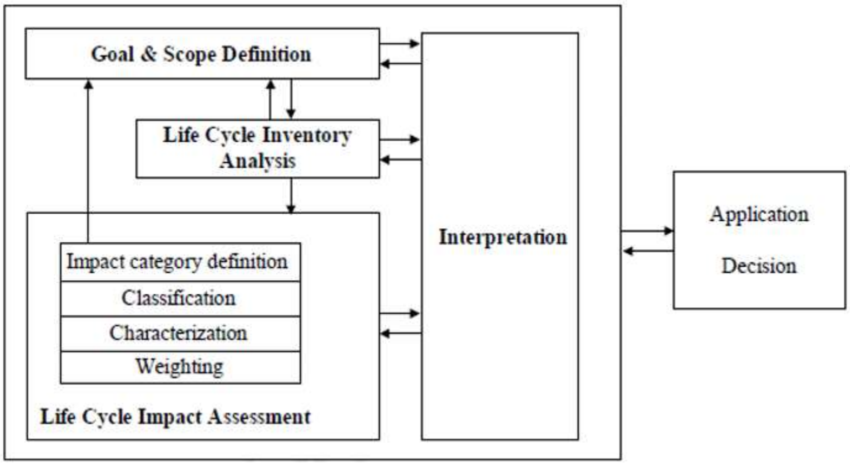


*Figure 1. LCA framework* [15].

If a full LCA is performed on the use phase of a dwelling, the following things should be taken into account: operational energy use (space heating and cooling, hot water consumption, building and user electricity, etc.), maintenance, embodied carbon, repair, replacement, refurbishment, and operational water use [16].

Differences in goal and scope, assumptions, and errors in input parameters make it challenging to compare different cases analyzed with an LCA model [17], [18], [19]. Due to the uniqueness of buildings there are a lot of different input parameters which makes comparison of different LCA challenging [20].

There are three types of LCAs: the process based LCA, the Economic Input-Output LCA, and the Hybrid LCA. The process-based LCA uses inputs and outputs for each process of a product. This is the recommended strategy by the ISO 14044:2006 standard. It is a detailed and accurate process, therefore the process-based approach needs a lot of data which could subsequently lead to high costs and time investment [16]. Data uncertainty and narrow boundary definition are also mentioned as disadvantages [21].

The Economic Input-Output (EIO) LCA is very suitable for large supply chains to quantify direct and indirect emissions. This method is very suitable to assess a geographical region of buildings. It is in general a faster method than the process-based LCA, if the databases are available. It is however also a less detailed approach than the process-based approach [16]. Kucukvar et al. [22] used it to analyze the emissions of the whole construction sector in the United States. The EIO-LCA method is only suitable for ex-post measurement since economic data is used. Therefore, this method is unfit for the prediction of the emissions during the use phase of a building.

The hybrid LCA is a combination of the process-based and the EIO-LCA. The hybrid model combines the other two methods so the completeness or accuracy in comparison to the other approaches is debatable [16]. This method combines data from both methods which would assure a more complete assessment. However, the variation in methodology in different cases makes comparison harder [21]. Since the hybrid LCA, like the EIO-LCA, uses economic data, this method is unfit for the purpose of this study.

### 2.1.2 Dynamic life cycle assessment

Traditional LCAs are used to assess the environmental impact of a building but do not take the time variation into account. The life span of a building is quite long, in literature varying between 40-100 years [23], and therefore the dynamic LCA aims to respond to this [15]. Dynamic LCA can be used to track potential changes, e.g. refurbishment, over a longer period. However, taking the time perspective into account also increases the complexity of the model. Moreover, the uncertainty of these assumptions should also be taken into account [24].

### 2.1.3 Life Cycle Energy Assessment

The Life Cycle Energy Assessment (LCEA) is a simpler version of the LCA which only focuses on energy to give insight in the different phases throughout the life cycle of a product. When LCEA is used it is important to specify whether primary or secondary energy is used, primary energy could be coal and secondary energy could be electricity [25].

According to Chau et al. [25], the operational phase can be analyzed in three ways. The first method is to use the actual measured energy consumption, so an ex-post measurement. The second method uses energy databases with building and location specific benchmark data to estimate the operational energy. So, this second method can predict average energy use. The third method uses simulation methods to estimate the operating energy. The third method of LCEA uses two simulation methods; the steady state model and the dynamic model of which the dynamic model takes the time variant of heating and cooling into account. Dynamic models are more complex. Both simulation methods are very sensitive to assumptions for the factors in the model [25].

Bribian et al. [20] use a simplified LCA method which only includes the operational energy in the use stage of a building. Meaning that maintenance, repair and replacement, and refurbishment are excluded. However, Martinez-Rocamora et al. [26] show the importance of the maintenance phase on the ecological footprint, especially cleaning activities.

A list of assumptions and uncertainties for the LCEA performed by Atmaca et al. [27] shows several relevant ones for the use phase of buildings. Like the buildings lifetime is assumed to be 50 years, the energy mix is constant over that 50 years, future price changes which influence energy consumption are not taken into account, inhabitant behavior, heating and cooling comfort is assumed to be constant [27].

### 2.1.4 Life cycle carbon emissions assessment

The Life Cycle Carbon Emission Assessment ($LCCO_2A$) or carbon footprint analysis takes all carbon emission equivalents into account over the life cycle of a building. This can basically be presented a sum of the $CO_2$ emissions of each phase of the life cycle of a building. So, $LCCO_2A$ is a subset of a full LCA which takes only the $CO_2$ emissions into account [16].

### 2.1.5 Building information modelling – life cycle assessment

Building Information Modelling (BIM) is a virtual 3D building model. The integration of BIM and LCA could in theory overcome the barrier of data acquisition [28]. The BIM software is not yet well integrated with sustainability databases so this will need significant time effort to import the information needed. There are several BIM-LCA integration tools which are suitable for the design stage, but Bueno et al. argue that for a full LCA a big software program like GaBi is needed [29].

### 2.1.6 Energy performance coefficient

Energy certification emerged in the early 1990s as a method to reduce energy use and subsequently $CO_2$ emissions. In 2002 the European Union introduced a regulatory instrument on energy performance of buildings. The instrument must include: an overall Energy Performance Index (EPI), meaning energy consumption and $CO_2$ emissions per unit (square meter) of conditioned area. Minimum efficiency requirement or a maximum EPI to improve performance. The label is based on a score from A-G to achieve a grading of buildings. This should relate to energy regulations, existing buildings stock and the zero-energy buildings [30], [31].

In the Netherlands this has resulted in the 'Energie Prestatie Coëfficiënt' (EPC), which calculates the building-related energy use. How the EPC of buildings is calculated is stated in the building decree. The EPC calculation takes the sum of the energy use by: space heating and cooling, humidification, fans (mechanical ventilation), lighting, hot water heating, and subtracts the self-generated energy. The losses and efficiencies of the installations and distribution systems are taken into account and compensated for [32].

The energy used for cooking and white and brown goods are excluded in this calculation because these are not building-related [33]. The EPC calculation assumes fixed: temperature settings, demand for hot water, lightning, and ventilation flow rates. These fixed values are based on standard use of building [33].

The EPC value corrects for the size of the dwelling. So, if a larger dwelling consumes more energy because of the size but has the same thermal quality as a smaller dwelling, the larger dwelling is not penalized. Thus the EPC value in this case could be the same [33].

The EPC is an instrument which has the goal of reducing the building-related energy consumption. Guerra Santin [33] shows that there is a statistically significant difference between dwellings built before the introduction of the EPC and after, this indicates that the EPC helped to reduce energy consumption in residential buildings.

Every building designer in the Netherlands is obligated to calculate the EPC score during the design stage, the EPC score is a dimensionless number. A building with a lower EPC is expected to use less energy than a building with a higher EPC. Since 2015 the EPC-score needs to be below 0,4 for domestic buildings in the Netherlands [34].

Guerra Santin [33] states that the difference in actual energy use in dwellings with various EPC scores is not statistically significant. Majcen et al. [35] found that the actual gas consumption is lower than the theoretical gas consumption in Dutch residential buildings. They show that residential buildings with a bad energy label consume much less energy than the label predicts. Energy-efficient buildings, on the contrary, consume more energy than predicted [35].

Since in the Netherlands building designers are obligated to obtain an EPC-score, using this score to predict the building specific energy use is a suitable method. However, the actual energy use during the use phase is different than the EPC-score indicates [33], [35].

### 2.1.7 Overview of approaches

This section presents an overview of the approaches described here above. Table 1 presents the suitability of the methods used in literature to predict the actual energy use during the use phase of a dwelling.

*Table 1.Overview of approaches used in literature to predict the actual use phase emissions of a dwelling.*

| Method | Ex-ante/Ex-post | Input data | Suitable |
|---|---|---|---|
| LCA process based | Ex-ante | Building specific parameters | No, time-consuming [16] |
| LCA EIO | Ex-post | Economic data | No, cannot be used to predict [22] |
| LCA hybrid | Ex-post | Building specific parameters and economic data | No, cannot be used to predict [16] |
| Dynamic LCA | Ex-ante | Building specific parameters | No, time-consuming [15] |
| LCEA 1 | Ex-post | Measured energy use | No, cannot be used to predict [25] |
| LCEA 2 | Ex-post | Energy databases and benchmark data | No, cannot be used to predict [25] |
| LCEA 3 | Ex-ante | Simulated data | No, time-consuming [25] |
| $LCCO_2A$ | Ex-ante | Building specific parameters | No, time-consuming [16] |
| BIM LCA | Ex-ante | Building specific parameters | No, not sufficiently developed [29] |
| EPC | Ex-ante | Building specific parameters | No, inaccurate [33], [35] |
| LCA (Life Cycle Assessment), EIO (Economic Input Output), LCEA (Life Cycle Energy Assessment), $LCCO_2A$ (Life Cycle $CO_2$ Assessment), BIM (Building Information Modelling), EPC (Energy Performance Certificate) | | | |

From Table 1 it becomes apparent that methods used in literature are not suitable for a contractor with the purpose of predicting the actual energy use during the use phase of a building in a time-efficient way.

## 2.2 Essential parameters in predicting use phase emission calculations

The EPC-score is not really accurate in predicting the actual energy use as stated before, there is a mismatch in actual and theoretical energy use. Therefore, this section presents more parameters, identified in previous research, which are essential to take into account in use phase emission calculation. This is split into two categories: building specific parameters, and other parameters.

### 2.2.1 Non-building specific parameters

There are several non-building specific parameters which are identified as important in use phase emission calculations. Ownership of the house and salary of the inhabitants are identified by Majcen et al. [36] as important predictors of actual gas use. Income is also mentioned by Guerra Santin [37] as an important predictor next to home amenities, family size and composition. Also Sardianou [38] shows that family size and annual income are influential parameters. Sardianou [38] also mentions age of the inhabitants and rate of occupancy as influential parameters. Gosselin et al. [39] identified occupant behavior as the parameter which caused the most variability between dwellings. In the regression analysis, opening windows in winter or using electrical appliances are most influential on the energy balance in apartments in Canada [39]. Satre-Meloy [40] shows that appliances and occupant behavior are associated with increased electricity usage.

Heesen et al. [41] researched consumer behavior in energy efficient homes in Germany to identify the usefulness of energy performance ratings as benchmark. They show that a few outliers in the dataset influence the actual energy use and therefore the prediction of energy use is found to be troublesome. Room temperature is the variable which mostly influences the actual heating energy consumption.

### 2.2.2 Building specific parameters

Building specific parameters are also important in emission calculations. Majcen et al. [36] identified floor area and value of the house as important parameters to predict the actual gas use. Heating area, building type and number of rooms are influential building specific parameters identified by Guerra Santin [37]. Several studies show that dwelling size is an important parameter in actual energy use [38], [40], [42]. The number of rooms , energy source, and building type are also highly influencing factors in predicting electricity use in Spanish households [42].

In a multiple regression analysis Carpino et al. [43] tested the following variables on influencing the dependent variable (heating demand) for Mediterranean residential buildings: geographical location, typology of external walls, windows, heating system, hot water heating system, gross surface divided by the heated volume, solar energy through windows,  and energy performance certificates. This study uses a sample of less than 200 houses which is small considering that 28 variables are tested. The coefficient of heat transfer is identified as the most important variable [43].

*Table 2. Building specific parameters identified in other studies as important predictors of the actual energy use during the use phase of dwellings.*

| **Building specific parameters** | **reference** |
|---|---|
| Floor area | [36] |
| Value of the house | [36] |
| Heating area | [37] |
| Building type | [37], [42] |
| Number of rooms | [37], [42] |
| Dwelling size | [38], [40], [42] |
| Energy source | [42] |
| Heat transfer coefficient | [43] |

# 3 Method

In this chapter the method of data collection, processing and analysis is elaborated on. The purpose of this chapter is to describe the method used to identify the important building-related parameters which predict actual energy use during the use phase of Dutch dwellings.

## 3.1 Study design

To identify predicting parameters of the actual energy use during the use phase of a Dutch dwelling a multiple linear regression analysis is used. The general forms of a simple linear regression model and a multiple linear regression model are shown in Equation 1 and Equation 2.

To perform a regression analysis, a dataset is needed with relevant building-specific parameters coupled to the actual energy use during the use phase of a Dutch dwelling.

A simple linear regression fits a line through the data which best describes a Dependent Variable (DV) using a constant and an Independent Variable (IV). The formula will be like Equation 1, where a and b are numbers.

*Equation 1. Simple linear regression model. Where DV = Dependent Variable, IV = Independent Variable, and a and b are numbers* [44].

$$DV = a * IV + b$$

A multiple linear regression attempts to model the relationship between multiple independent variables to predict a dependent variable. So, the formula will be of the same structure as Equation 2.

*Equation 2. Multiple linear regression model. Where DV = Dependent Variable, IV = Independent variable, and a and bn are numbers* [44].

$$DV = a + b_1 * IV_1 + b_2 * IV_2 + \cdots + b_n * IV_n$$

In SPSS there are 5 methods to perform a multiple linear regression analysis (enter, stepwise, remove, forward, and backward). *The enter method* forces all the independent variables in a model at once, without regards to the independent variables making a significant contribution to the model [44]. *The remove method* removes all independent variables in a single step (only relevant if user specifies multiple steps) [44]. *The Backward selection method* first enters all independent variables and then removes them one at a time based on a significance level, the least significant predictor which meets the exclusion criteria will be deleted [44]. *The forward method* only adds variables which make a significant prediction to the model, variables are entered in one at a time, starting with the most significant predictor which meets the inclusion criteria [44]. *The stepwise method* uses both the forward and backward regression. It starts with forward regression and adds the most significant independent variable of the model. Then a backward step is used to check if there are independent variable in the model that need to be excluded which is the case if one of the predictors of a previous step has become an insignificant predictor. This method continues until there are no longer any predictor variables that meet the criteria for entering the model or being removed from the model [45].

In this study the dependent variable to predict is: the actual energy use during the use phase of a Dutch dwelling. The aim of this study is to predict this dependent variable with building-specific independent variables according to the criteria.

Multiple linear regression analysis is used to create a model to predict the energy consumption in Dutch dwellings using building characteristics. In a first step, stepwise multiple linear regression is used to determine the importance of variables in the model. In the second step all categorical (nominal) variables are dichotomized and forced into the multiple regression model (using the enter method) to subtract coefficients of the variables, where 80% of the data is used to create the models and subtract the coefficients and the rest of the data will be used to cross-validate the model. This cross-validation step ensures that there is no overfitting of the sample.

## 3.2 Data collection

At BAM the data availability is limited, for one type of building the actual energy use was measured. Since this is only 1 type of building, the dataset of BAM is insufficient to use for a multiple regression analysis for identifying key parameters in predicting the actual energy use. Therefore, an alternative dataset was consulted in the form of the 'Woon Onderzoek Nederland' (WoON) dataset [1].

Some background information of the research performed by WoON:

*The study carried out by 'Binnenlandse Zaken en Koninkrijksrelaties' (BZK) and 'Centraal bureau voor de Statistiek' (CBS), WoON is a large-scale research with a lot of themes. Themes included: rental-property or owner-occupied property, relocation tendency, building type, recently moved, monthly rent, desired municipality, desired neighborhood, home satisfaction, satisfaction living environment, involvement in livability of the neighborhood, interest in private commissioning, interest in buying own rental property, and educational level. The respondents are a random sample of the Dutch population* [1].

*The CBS collected the data between August 2017 and April 2018. In total 110.000 persons were approached to participate in the study. The questions were amongst others about energy use, maintenance and mortgage. Respondents were asked to fill out an extensive survey, which was combined with actual energy and gas data from the distribution companies and data from CBS. Inhabitants of the Netherlands which were randomly selected got a letter to participate as respondent in the research. With a gift card of 5 euro's as incentive* [1].

## 3.3 Data processing

The dataset of WoON [1] contains over 900 different variables in the various themes described above. This study aims to identify key building-related parameters to predict the actual energy use during the operation phase of Dutch dwellings. These 900 variables contain building-related variables which are relevant for this study but also variables which are irrelevant (e.g. about mortgage and searching for social housing). Therefore, the irrelevant variables are deleted. Criteria which variables need to meet to be kept in the dataset are:

1. The variable is a building-specific parameter.
2. The variables in the dataset can be known before the building is occupied, so in the design or construction stage. This design information is available to a contractor and can be used to predict the actual energy use in the operation phase of a building.
3. The variables contain enough respondents, and that the variables contain objective answers.

Actual energy use, gas and electricity, are also kept as variables in the dataset. The total actual energy use is created as extra variable by adding the actual gas use to the actual electricity use. The total actual energy use is the dependent variable of which the prediction or calculation is the goal of this study.

Variables that are kept in the dataset are the building specific parameters (gross floor area, building type, number of rooms, etc.), and the heating systems, PV panels, energy efficiency measures, and EPC-score.

As described in section 2.1.6 the EPC-score corrects for dwelling size. So, to predict the actual energy use the EPC score needs to be multiplied with the corresponding Gross Floor Area (GFA). The variable which combines these two is in this study called: the theoretical total energy use.

## 3.4 Data analysis

The dataset of WoON 2018 [1] was analyzed in SPSS 23.

Regression analysis is a statistical tool to analyze the relationship between two or more variables. Basically, to test the influence of one or more independent variables on a dependent variable. The purpose of the regression analysis of the dataset in this study is analyzing and exploring which parameters are key in predicting the actual energy use during the operation phase of a building.

Stepwise multiple regression analysis was used to check which parameters are needed towards a prediction of the actual energy use. When the predicting variables are identified, the nominal variables are dichotomized (meaning that each category becomes a variable with two possible outcomes 0 (no) or 1 (yes)). This is necessary to be able to retrieve the coefficients and create the prediction models.

These requirements need to be taken into account for performing a multiple linear regression [46]:

1. The dependent variable should be measured on a continuous scale
2. There are two or more independent variables (continuous or categorical)
3. Independence of observations is assumed
4. The linear relationship between the dependent variables and each of the independent variables is assumed
5. The data shows homoscedasticity
6. The data must not show multicollinearity
7. There are no significant outliers, high leverage points or highly influential points
8. Residuals should be approximately normally distributed

If the $R^2$ value, which predicts the amount of variance predicted by the model, is bigger than 0.3 the model is considered a good fit of the data. Further analysis on the predictive value of the model will then be presented in the discussion chapter.

Only respondents living in buildings constructed in 2008 or later are included in the sample analysis. This is because in 2008 the EPC score was introduced in the Netherlands [47].

# 4 Results

This chapter presents the results as they were obtained. This chapter starts with the description of the dataset, followed by the multiple linear regression analysis and finishes with the relevant models and its coefficients.

## 4.1 Description of dataset

The variables used in the multiple regression analysis are shown in Table 3 together with their type of data. The independent variables first and the dependent variables are on the bottom of the table. The answering possibilities for the nominal variables are listed in Appendix A.

*Table 3. Independent and dependent variable labels with short notation and type of data.*

| **Variable Label** | **Type of data** |
|---|---|
| **Independent variables:** | |
| Building type | nominal |
| Number of floors | scale |
| Number of rooms | scale |
| Living room area ($m^2$) | scale |
| Number of habitable floors | scale |
| Energy saving measures (double glass, insulation, solar panels, new heater system, other, none) | nominal |
| Gross floor area ($m^2$) | scale |
| Construction year | scale |
| Energy label | nominal |
| Garage or carport | nominal |
| Building especially for elderly people | nominal |
| Living room on what floor | scale |
| Theoretical gas use ($m^3$) | scale |
| Theoretical electricity use (kWh) | scale |
| Theoretical total energy use (kWh) | scale |
| Heater type | nominal |
| Hot water heater type | nominal |
| **Dependent variables:** | |
| Electricity use (kWh) | scale |
| Gas use ($m^3$) | scale |
| Total energy use (kWh) | scale |

As described in section 3.3 outliers and respondents with a house older than construction year 2008 are excluded from the dataset. Meaning that all figures and calculations in this chapter only present buildings built after 2008. After the exclusion criteria 1746 respondents/dwellings were kept in the dataset.

## 4.2 Multiple linear regression

A multiple linear regression was performed to find predictors of the dependent variable: ‘actual energy use during the operation phase of a Dutch residential building’. The assumptions for the analysis are met and are described in Appendix B. The summary of the multiple regression analysis is presented in Table 4, with as dependent variable the total actual energy use.

*Table 4. Summary of multiple linear regression models[2].*

| Approximately 80% of the cases (SAMPLE) | Model | R | R Square | Adjusted R Square | Std. Error of the Estimate | Change Statistics | | | | | Durbin-Watson |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | R Square Change | F Change | df1 | df2 | Sig. F Change | |
| Selected | 1 | ,539[a] | ,291 | ,291 | 5828,261 | ,291 | 1298,139 | 6 | 18986 | ,000 | |
| | 2 | ,578[b] | ,334 | ,334 | 5648,272 | ,043 | 1230,301 | 1 | 18985 | ,000 | |
| | 3 | ,592[c] | ,351 | ,351 | 5576,840 | ,017 | 123,365 | 4 | 18981 | ,000 | |
| | 4 | ,594[d] | ,353 | ,352 | 5569,954 | ,002 | 7,709 | 7 | 18974 | ,000 | |
| | 5 | ,597[e] | ,356 | ,355 | 5556,996 | ,003 | 15,766 | 6 | 18968 | ,000 | |
| | 6 | ,597[f] | ,357 | ,356 | 5554,020 | ,001 | 11,166 | 2 | 18966 | ,000 | 1,956 |

Table 4 shows that the multiple linear regression models 2, 3, 4, 5 and 6 are considered a good fit. The models are created that each new model adds an extra predicting variable in comparison to the previous model, e.g. model 2 has 2 variables and model 1 has 1 variable. The Sig. F Change column shows that all the models are statistically significant (values are <0.05).

The most relevant models are model 3 and 6. The R Square change column shows that model 3 is 1.7% (0.017*100) better explaining the variance than model 2. Model 4 explains 0.2% more variance than model 3, model 5 0.3% more than model 4, and model 6 0.1% more than model 5. So, the last 3 variables added, resulting in model 6, together explain 0.6% of the variance. The exact coefficients of the variables in the models are presented in Appendix C.

The equations of model 3 and model 6 are presented in Equation 3 and Equation 4 respectively. Model 3 (Equation 3) includes a constant and the variables: building type, theoretical total energy use, and the number of rooms (in 5 classes). Since the building type and the number of rooms (in 5 classes) are nominal variables these are shown in the equation as dichotomous variables.

*Equation 3. Multiple linear regression model 3. Where all variables are dichotomized except for theoretical total energy use.*

$$\begin{aligned} 8810 + 1945 * (terraced\ house) + 4080 * (semidetached\ house) + 9418 * (detached\ house) \\ + 2875 * (farm) + 2599 * (home\ with\ store) + 3450 \\ * (residential\ unit\ with\ cummunal\ facilities) + 0.428 \\ * (theoretical\ total\ energy\ use) - 173 * (rooms(3)) + 1054 * (rooms(4)) \\ + 2031 * (rooms(5)) + 3543 * (rooms(6+)) \end{aligned}$$

Model 6 (Equation 4) includes a constant and the variables: building type, theoretical energy use, the number of rooms (in 5 classes), hot water heater type, GFA (7 classes), and garage or carport. All these variables, except theoretical total energy use, are nominal variables. These nominal variables are shown in the equation as dichotomous variables.

[2] 80% of sample data (randomly selected) is used. Dependent variable: total energy use. A: (constant) and building type), b: (constant, building type and theoretical energy use), c: (constant, building type, theoretical energy use and number of rooms), d:(constant, building type, theoretical energy use, number of rooms, and hot water heater type), e: (constant, building type, theoretical energy use, number of rooms (5 classes), hot water heater type, and gross floor area(7 classes)), f: (constant, building type, theoretical energy use, number of rooms (5 classes), hot water heater type, gross floor area(7 classes), and garage or carport), g: (dependent variable: total energy use).

*Equation 4. Multiple linear regression model 6. Where all variables are dichotomized except for theoretical total energy use.*

$$8875 + 1984 * (terraced\ house) + 3579 * (semidetached\ house) + 8512 * (detached\ house) + 1695 * (farm) + 1772 * (home\ with\ store) + 3976 * (residential\ unit\ with\ communal\ facilities) + 0.377 * (theoretical\ total\ energy\ use) - 174 * \big(rooms(3)\big) + 903 * \big(rooms(4)\big) + 1678 * \big(rooms(5)\big) + 2952 * \big(rooms(6+)\big) - 337 * (combined\ CV\ and\ solar\ boiler) + 872 * (gas\ water\ heater) + 670 * (electrical\ boiler) - 170 * (heatpump\ water\ heater) + 1318 * (block\ or\ district\ water\ heating) - 64 * (district\ water\ heating) + 190 * (no\ water\ heating) - 313 * \big(GFA(50-69m2)\big) - 107 * \big(GFA(70-89m2)\big) + 265 * \big(GFA(90-119m2)\big) + 670 * \big(GFA(120-149m2)\big) + 1176 * \big(GFA(150-199m2)\big) + 2208 * \big(GFA(200+m2)\big) + 488 * (garage) + 868 * (carport)$$

The cross-validation of model 3 and model 6 are shown in Table 5 and Table 6 respectively.

*Table 5. Pearson correlations between model 3 (predictedmodel3) and the actual total energy use (Totale energieverbruik (kWh)).*

| Approximately 80% of the cases (SAMPLE) | | | predictedmodel3 | Totale energieverbruik (kWh) |
|---|---|---|---|---|
| Not Selected | predictedmodel3 | Pearson Correlation | 1 | ,590** |
| | | Sig. (2-tailed) | | ,000 |
| | | N | 4780 | 4718 |
| | Totale energieverbruik (kWh) | Pearson Correlation | ,590** | 1 |
| | | Sig. (2-tailed) | ,000 | |
| | | N | 4718 | 13229 |
| Selected | predictedmodel3 | Pearson Correlation | 1 | ,592** |
| | | Sig. (2-tailed) | | ,000 |
| | | N | 19185 | 18993 |
| | Totale energieverbruik (kWh) | Pearson Correlation | ,592** | 1 |
| | | Sig. (2-tailed) | ,000 | |
| | | N | 18993 | 52919 |

*Table 6. Pearson correlation between model 6 (predictedmodel6) and the actual total energy use (Totale energieverbruik (kWh)).*

| Approximately 80% of the cases (SAMPLE) | | | predictedmodel6 | Totale energieverbruik (kWh) |
|---|---|---|---|---|
| Not Selected | predictedmodel6 | Pearson Correlation | 1 | ,598** |
| | | Sig. (2-tailed) | | ,000 |
| | | N | 4780 | 4718 |
| | Totale energieverbruik (kWh) | Pearson Correlation | ,598** | 1 |
| | | Sig. (2-tailed) | ,000 | |
| | | N | 4718 | 13229 |
| Selected | predictedmodel6 | Pearson Correlation | 1 | ,597** |
| | | Sig. (2-tailed) | | ,000 |
| | | N | 19185 | 18993 |
| | Totale energieverbruik (kWh) | Pearson Correlation | ,597** | 1 |
| | | Sig. (2-tailed) | ,000 | |
| | | N | 18993 | 52919 |

Table 5 and Table 6 show the Pearson correlation between the multiple linear regression models 3 and 6 respectively and the actual total energy use. These tables are split in two categories (Selected) 80% of the data which was used to create the prediction model and (Not Selected) 20% of the data which was used to validate the prediction models. The correlation is for both categories (Selected and Non-Selected) and in both models approximately the same (0.590 and 0.592 for model 3) (0.598 and 0.596 for model 6), implying that there is no overfitting of the data.

# 5 Case study/real application of method

To show the relevance of these results an advice is presented to BAM. The intention of BAM is to improve its emission reporting. BAM as a company is split in three business lines with their own main activities. These three business lines are split in business units per country. This case study focuses on improving the calculation method of BAM by adding one business unit (BAM Wonen). This chapter will in short present the current method, a new method with existing data, and a suggestion for implementation of the model for next year.

The current method used for emission calculation and reporting uses one business unit of BAM, BAM Construct UK. Data of BAM Construct UK is used and combined with the UK energy performance certificate emission standards (standard emissions for certain energy performance certificate score) or reference data of CIBSE to calculate/estimate the emissions of this business unit. These emissions are then coupled to revenue and subsequently extrapolated over revenue of the business line: BAM Construction M&E services (BAM CME). This method is used for 2017 and 2018.

The new method with existing data is performed as follows. To update the data of 2018 extra data was collected from BAM Wonen (business unit which focuses on Dutch residential buildings and represents approximately 10% of the revenue of BAM CME). A list with the number of buildings built in 2018 with the EPC score was used to estimate the use phase emissions of BAM Wonen. The actual gross floor area was not available data, so the gross floor area was estimated to be 114m$^2$, this is the average dwelling built in 2018 in the Netherlands [48]. To estimate the total emissions of BAM CME, the emissions calculated for BAM Wonen and BAM Construct are coupled to revenue and subsequently extrapolated over revenue of BAM CME. Figure 2 gives an overview of the three situations.

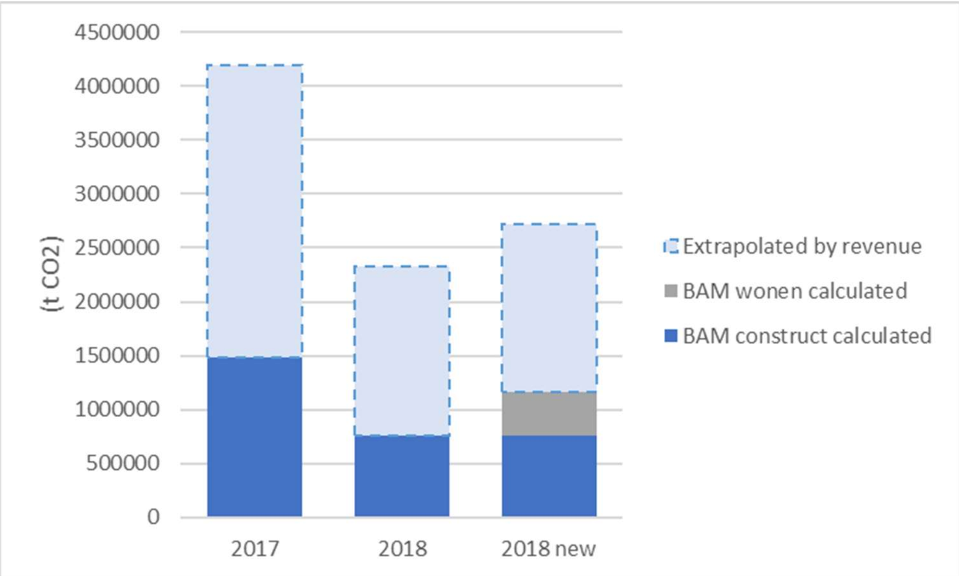


*Figure 2. CO2 emissions of the 'Use of sold products' of BAM CME calculated using 3 different methods.*

Figure 2 shows a big difference between 2017 and 2018 which can be explained by the number of finished projects in both years. In 2017 the number of finished projects was higher, while the revenue was approximately the same, which means that the emissions reported in that year are higher since only finished projects are emission wise reported.

The results chapter shows 2 different multiple regression models which are statistically significant. At BAM it is an important criterium that the information is easily obtained since it is a commercial company. Therefore, the models were also judged on the added value of collecting an extra parameter. The cumulative predicting value of parameters are shown in Figure 3. The predicting value is composed of the adjusted $R^2$ multiplied by 100 to give a clear, understandable, overview.

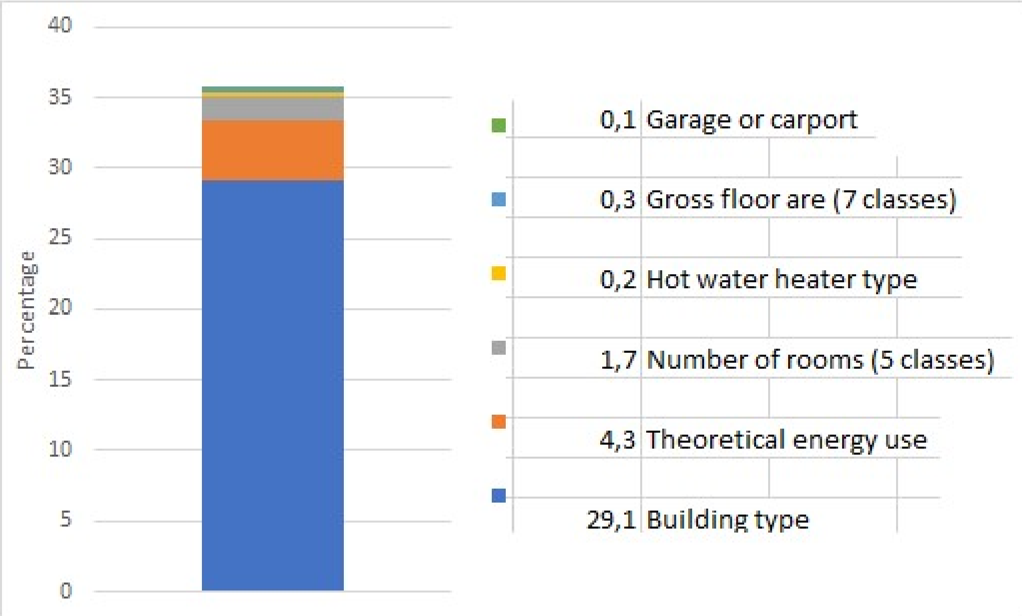


*Figure 3. Cumulative predicting value of parameters of multiple regression model.*

## 5.1 Application

Currently the data availability relevant for emission calculations is limited within BAM. There is not yet a central system which keeps important data. There are two main possibilities to start collecting data to improve the $CO_2$ emission calculations. 1) The first option is to use the Customer Relationship Management (CRM) system. This system is a database in which all projects are noted down with some obligatory fields to fill out (like project number and revenue). This system has the possibility to expand with extra obligatory or non-obligatory fields to collect the extra data needed. This would require some extra work for the people filling the form. An advantage to this option is that in coming year this system will be introduced/implemented within all business units of BAM (which also involves training to work with the system). 2) The second option is to use BIM. BIM is quite an elaborate tool which could contain all these parameters. However, within BAM the coverage among dwellings is quite low. To expand this coverage would require quite some work and training in BIM.

To conclude the case study, for 2018 more data is used to improve the accuracy of the emission calculation. Next year extra data can be collected using the CRM system or BIM. BIM requires training and knowledge so it will take some time to implement and is perhaps unrealistic for next year. The CRM system can be updated with obligatory fields to fill out. BAM Wonen represents 10% of the revenue of BAM CME. If regression model 3 which uses 4 parameters (building type, theoretical energy use, and number of rooms (5 classes) is implemented to predict the actual energy use. The calculation of BAM Wonen will improve up to 35% accuracy. BAM Wonen represents 10% of the revenue of BAM so the overall predictive value will improve almost 3.5%. If in the future BIM will be standard for all buildings this system could be used to automate these calculations.

# 6 Discussion

This chapter starts with the interpretation of the results and comparison with other studies followed by the limitations of this study and possible future research.

## 6.1 Interpretation of results

This study shows that 35.7% of the variance in actual total energy use can be explained by a linear model with 7 parameters, if 4 parameters are used 35.1% of the variance can be explained. The main contribution of this study is that it uses few and easily available building specific parameters to predict the actual energy use during the operation phase of a Dutch dwelling.

The results of this study can be used to predict the actual total energy use of Dutch dwellings built after 2008. In the best predicting models of the actual total energy use, the last three variables which are added to the model together just explain an extra 0.6% of the variance, which means they don't explain a lot of the data variance but are statistically significant predictors. Therefore, the added value of the last three variables (hot water heater type, GFA in 7 classes, and garage or carport) can be questioned.

This study shows that the theoretical energy use by itself is not sufficient in predicting the actual energy use, which is similar in other studies [33], [35]. The most important parameter in predicting the actual energy use is, as stated above, the building type which is in agreement with a study by Majcen et al. [35]. The second most important predictor in this study is the theoretical total energy use which is the EPC score and the GFA combined. The GFA is by others identified as an important parameter, some studies use another term than GFA (e.g. dwelling size or heating area) [37], [38], [40], [42]. There is a slight difference, this study combines the GFA with the EPC score instead of GFA as a single variable. The GFA is as a single variable not identified as an important variable, this can partly be caused by the correlation between the theoretical energy use and the GFA of which the theoretical energy use is a better predictor so chosen over GFA. The number of rooms as a predicting parameter is in a previous study shown to be an important parameter in predicting electricity and therefore energy use [42]. So, the variables which are used in the regression models are in line with previous studies.

## 6.2 Limitations

The method used in this study is a stepwise multiple linear regression. An issue with this method is the risk of overfitting the model. In this study 80% of the sample is used to create the regression models and 20% of the data is used to validate the model to show that overfitting is not an issue [49].

Carpino et al. [43] identify the coefficient of heat transfer as the most important variable, this variable has not been studied because this parameter was not available in the dataset.

This study is limited to Dutch dwellings, climate and climate-related variables are beyond the scope of this thesis. However, the approach taken in this study could be useful for future research in other countries with different climate conditions. This study focuses on buildings built after 2008 since the EPC score is obligatory since 2008 in the Netherlands. The goal of this study is to predict the actual energy use before the building is occupied, in older buildings old energy use data could be used to predict actual energy use. If fewer years would have been taken into account the sample size would have been smaller so the results would be less valid.

The energy labels used in this study are categorical variables A-G instead of the exacter (scale) EPC value. These energy labels are combined with the GFA to get the theoretical total energy use. So, this theoretical total energy use which is a scale variable has a rounding error. In this dataset an exacter value like EPC was not available. The WoON dataset has several building specific parameters available and has given the opportunity to use a big dataset, which is appropriate for the method used in this study.

This study uses stepwise multiple linear regression as a method to identify key parameters in predicting the total actual energy use of a Dutch residential building. This is an approach often used in exploratory studies. This has some risks like: that the adjusted $R^2$ value could be too high, and the regression coefficients could be biased [44].

# 7 Conclusions

The main research question of this study is: *How can the energy related emissions during the use phase of a Dutch dwelling be predicted using building specific data?*

To answer the main research question the sub questions are answered in this section followed by an answer to the main research question.

This study shows that there are several existing approaches available for the $CO_2$ emission calculations of the use phase of dwellings. The approaches are: LCA, dynamic LCA, LCEA, $LCCO_2A$, BIM LCA and EPC. All approaches are either time-consuming due to data-need, not predicting (ex-post) or inaccurate for predicting the actual energy use of a dwelling. This study used stepwise multiple linear regression to create a new prediction model to calculate the energy use in the use phase of Dutch dwellings. The key parameters in predicting the actual energy use of a residential building are: the building type, the theoretical total energy use (EPC score combined with GFA) and the number of rooms. The multiple linear regression model predicts up to 35% of the variance of the data. Time-efficiency and data-need are important parameters for the implementation of a new calculation method at a contractor. This study shows that a relatively simple calculation with 4 variables can be used to predict 35% of the variance in the total actual energy use. Therefore, the model with the parameters described above is suitable for a quick implementation.

The model presented in this study uses 4 building specific parameters and can be used to predict 35% of the variance in the actual energy use during the use phase of Dutch dwellings built after 2007. This prediction of energy use can be used to calculate the energy related emissions.

# 8 Recommendations and future work

On purpose this study focused on residential buildings and not on civil engineering structures because very little is known about emissions during the use phase of these structures. But this can be an interesting topic for future research.

With the building specific parameters used in this study almost 35% of the variance in actual energy use can be explained. This study focused on building specific parameters, because these parameters can be influenced by a contractor, while perhaps also other parameters are of influence on the actual energy use. Other possible parameters are: income, home amenities, family size and composition, and occupant behavior. This is beyond the scope of this study but could be interesting for future studies as well. Where both the prediction of the actual energy use is interesting, but influencing occupant behavior to reduce the actual energy use could also be an interesting topic for future studies.

# References


[1] BZK and CBS, "WoON." 2018.

[2] A. Core writing team; Pachauri, R, K; Meyer, L, *IPCC, 2014: Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC)*, vol. 40, no. 1. 2014.

[3] UN Environment and International Energy Agency, *Towards a zero-emission, efficient, and resilient buildings and construction sector. Global Status Report 2017.* 2017.

[4] R. Santos, A. A. Costa, J. D. Silvestre, and L. Pyl, "Integration of LCA and LCC analysis within a BIM-based environment," *Autom. Constr.*, vol. 103, no. February, pp. 127–149, 2019.

[5] B. Comyns, "Climate change reporting and multinational companies: Insights from institutional theory and international business," *Account. Forum*, vol. 42, no. 1, pp. 65–77, 2018.

[6] "About Us | Greenhouse Gas Protocol." [Online]. Available: http://ghgprotocol.org/about-us. [Accessed: 07-May-2019].

[7] G. Harangozo and C. Szigeti, "Corporate carbon footprint analysis in practice – With a special focus on validity and reliability issues," *J. Clean. Prod.*, vol. 167, pp. 1177–1183, 2018.

[8] (World Resources Institute) and (World Business Council for Sustainable Development), *Corporate Value Chain (Scope 3) Standard | Greenhouse Gas Protocol*. 2011.

[9] G. Hertwich, Edgar and R. Wood, "The growing importance of scope 3 greenhouse gas emissions from industry," *Environ. Res. Lett.*, vol. 13, no. 10, 2018.

[10] O. Ortiz-Rodríguez, F. Castells, and G. Sonnemann, "Life cycle assessment of two dwellings: One in Spain, a developed country, and one in Colombia, a country under development," *Sci. Total Environ.*, vol. 408, no. 12, pp. 2435–2443, 2010.

[11] N. Soares *et al.*, "A review on current advances in the energy and environmental performance of buildings towards a more sustainable built environment," *Renew. Sustain. Energy Rev.*, vol. 77, no. February, pp. 845–860, 2017.

[12] A. F. Abd Rashid and S. Yusoff, "A review of life cycle assessment method for building industry," *Renew. Sustain. Energy Rev.*, vol. 45, pp. 244–248, 2015.

[13] X. Gong, Z. Nie, Z. Wang, S. Cui, F. Gao, and T. Zuo, "Life cycle energy consumption and carbon dioxide emission of residential building designs in Beijing: A comparative study," *J. Ind. Ecol.*, vol. 16, no. 4, pp. 576–587, 2012.

[14] M. Buyle, J. Braet, and A. Audenaert, "Life cycle assessment in the construction sector: A review," *Renew. Sustain. Energy Rev.*, vol. 26, pp. 379–388, 2013.

[15] S. Su, X. Li, Y. Zhu, and B. Lin, *Dynamic LCA framework for environmental impact assessment of buildings*, vol. 149. Elsevier B.V., 2017.

[16] A. E. Fenner *et al.*, "The carbon footprint of buildings: A review of methodologies and applications," *Renew. Sustain. Energy Rev.*, vol. 94, no. July, pp. 1142–1152, 2018.

[17] X. Zhang, R. Zheng, and F. Wang, "Uncertainty in the life cycle assessment of building emissions: A comparative case study of stochastic approaches," *Build. Environ.*, vol. 147, pp. 121–131, 2019.

[18] L. Bragança, R. Mateus, and H. Koukkari, "Building sustainability assessment," *Sustainability*,

vol. 2, no. 7, pp. 2010–2023, 2010.

[19] B. Rossi, A. F. Marique, M. Glaumann, and S. Reiter, “Life-cycle assessment of residential buildings in three different European locations, basic tool,” *Build. Environ.*, vol. 51, pp. 395–401, 2012.

[20] I. Zabalza Bribián, A. Aranda Usón, and S. Scarpellini, “Life cycle assessment in buildings: State-of-the-art and simplified LCA methodology as a complement for building certification,” *Build. Environ.*, vol. 44, no. 12, pp. 2510–2520, 2009.

[21] R. Azari and N. Abbasabadi, “Embodied energy of buildings: A review of data, methods, challenges, and research trends,” *Energy Build.*, vol. 168, pp. 225–235, 2018.

[22] M. Kucukvar and O. Tatari, “Towards a triple bottom-line sustainability assessment of the U.S. construction industry,” *Int. J. Life Cycle Assess.*, vol. 18, no. 5, pp. 958–972, 2013.

[23] H. Islam, M. Jollands, and S. Setunge, “Life cycle assessment and life cycle cost implication of residential buildings - A review,” *Renew. Sustain. Energy Rev.*, vol. 42, pp. 129–140, 2015.

[24] C. K. Anand and B. Amor, “Recent developments, future challenges and new research directions in LCA of buildings: A critical review,” *Renew. Sustain. Energy Rev.*, vol. 67, pp. 408–416, 2017.

[25] C. K. Chau, T. M. Leung, and W. Y. Ng, “A review on life cycle assessment, life cycle energy assessment and life cycle carbon emissions assessment on buildings,” *Appl. Energy*, vol. 143, no. 1, pp. 395–413, 2015.

[26] A. Martínez-Rocamora, J. Solís-Guzmán, and M. Marrero, “Toward the ecological footprint of the use and maintenance phase of buildings: Utility consumption and cleaning tasks,” *Ecol. Indic.*, vol. 69, pp. 66–77, 2016.

[27] A. Atmaca and N. Atmaca, “Life cycle energy (LCEA) and carbon dioxide emissions (LCCO<inf>2</inf>A) assessment of two residential buildings in Gaziantep, Turkey,” *Energy Build.*, vol. 102, pp. 417–431, 2015.

[28] B. Soust-Verdaguer, C. Llatas, and A. García-Martínez, “Critical review of BIM-based LCA method to buildings,” *Energy Build.*, vol. 136, pp. 110–120, 2017.

[29] C. Bueno and M. M. Fabricio, “Comparative analysis between a complete LCA study and results from a BIM-LCA plug-in,” *Autom. Constr.*, vol. 90, no. November 2017, pp. 188–200, 2018.

[30] “Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings (EPBD),” *Off. J. Eur. Union*, pp. 65–71, 2002.

[31] L. Pérez-Lombard, J. Ortiz, R. González, and I. R. Maestre, “A review of benchmarking, rating and labelling concepts within the framework of building energy certification schemes,” *Energy Build.*, vol. 41, no. 3, pp. 272–278, 2009.

[32] C. Nobel, K, J and J. Haartsen, T, “Leeswijzer NEN 7120,” no. september, pp. 1–59, 2011.

[33] O. Guerra-Santin, *Actual energy consumption in dwellings. The effect of energy performance regulations and occupant behaviour.* 2010.

[34] (Rijksdienst voor Ondernemend Nederland), “Energieprestatie (EPC) |RVO.nl.” [Online]. Available: https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-en-regels-gebouwen/nieuwbouw/energieprestatie-epc. [Accessed: 03-Jun-2019].

[35] D. Majcen, L. C. M. Itard, and H. Visscher, “Theoretical vs. actual energy consumption of

labelled dwellings in the Netherlands: Discrepancies and policy implications," *Energy Policy*, vol. 54, pp. 125–136, 2013.

[36] D. Majcen, L. Itard, and H. Visscher, "Actual and theoretical gas consumption in Dutch dwellings: What causes the differences?," *Energy Policy*, vol. 61, pp. 460–471, 2013.

[37] O. G. Santin, "Behavioural patterns and user profiles related to energy consumption for heating," *Energy Build.*, vol. 43, no. 10, pp. 2662–2672, 2011.

[38] E. Sardianou, "Estimating space heating determinants: An analysis of Greek households," *Energy Build.*, vol. 40, no. 6, pp. 1084–1093, 2008.

[39] J. Rouleau, L. Gosselin, and P. Blanchet, "Understanding energy consumption in high-performance social housing buildings: A case study from Canada," *Energy*, vol. 145, pp. 677–690, 2018.

[40] A. Satre-Meloy, "Investigating structural and occupant drivers of annual residential electricity consumption using regularization in regression models," *Energy*, vol. 174, pp. 148–168, 2019.

[41] F. Heesen and R. Madlener, "Consumer behavior in energy-efficient homes: The limited merits of energy performance ratings as benchmarks," *Energy Build.*, vol. 172, pp. 405–413, 2018.

[42] M. C. Sánchez-Sellero and P. Sánchez-Sellero, "Variables determining total and electrical expenditure in Spanish households," *Sustain. Cities Soc.*, vol. 48, no. July 2018, p. 101535, 2019.

[43] C. Carpino, R. Bruno, and N. Arcuri, "Statistical analysis of the heating demand in residential buildings located in Mediterranean climate and proposals for refurbishment," *Energy Procedia*, vol. 133, pp. 16–27, 2017.

[44] A. Field, *Discovering Statistics Using IBM SPSS*, 4th ed. SAGE Publications Ltd, 2013.

[45] "Linear Regression Variable Selection Methods." [Online]. Available: https://www.ibm.com/support/knowledgecenter/en/SSLVMB_24.0.0/spss/base/linear_regression_methods.html. [Accessed: 18-Aug-2019].

[46] Laerd statistics, "How to perform a Multiple Regression Analysis in SPSS Statistics | Laerd Statistics." [Online]. Available: https://statistics.laerd.com/spss-tutorials/multiple-regression-using-spss-statistics.php. [Accessed: 11-Jun-2019].

[47] "Energielabel gebouwen utiliteit | bedrijfspand | RVO.nl." [Online]. Available: https://www.rvo.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-en-regels-gebouwen/bestaande-bouw/energielabel-utiliteitsgebouwen. [Accessed: 17-Jul-2019].

[48] CBS, "StatLine - Voorraad woningen; gemiddeld oppervlak; woningtype, bouwjaarklasse, regio." [Online]. Available: https://opendata.cbs.nl/statline/#/CBS/nl/dataset/82550NED/table?ts=1562574673270. [Accessed: 18-Jul-2019].

[49] R. R. Picard and R. D. Cook, "Cross-Validation of Regression Models," *J. Am. Stat. Assoc.*, vol. 79, no. 387, pp. 575–583, Sep. 1984.

[50] G. D. Garson, "Testing statistical assumptions: Blue Book Series," pp. 1–54, 2012.

# Appendices

This chapter presents additional information to provide the reader with extra information if required.

## Appendix A

Nominal variables in dataset with the answering options.

*Table 7. Answering possibilities nominal variables.*

| Label | Answering possibilities |
| --- | --- |
| Building type | Flat (apartment), terraced house, semi-detached house, detached house, farm, house with extra function (store), house with communal facilities, other, no answer |
| Energy saving measures (double glass, insulation, solar panels, new heater system, other, none) | Yes or no (for all individual questions), no answer |
| Energy label | A, B, C, D, E, F, G |
| Garage or carport | garage, carport, neither, refuses to answer |
| Building especially for elderly people | yes, no, no answer |
| Heater type | Central-heating boiler, wood-fired heater, pellet-heater, gas-heater, heat-pump, block or district heating, city district heating, other, no answer |
| Hot water heater type | Central-heating boiler (CV), combination of central-heating and solar boiler (CV and solar boiler), gas water heater, electrical boiler, heat-pump water heater, block or district water heater, city district water heater, other, no answer |

## Appendix B

Assumption testing

Assumption 1 is met since the actual total energy use is created from the total actual gas use and total actual electricity use which are both measured on a continuous scale. Therefore, the total actual energy use is also measured on a continuous scale.

Assumption 2 is tested by looking at the table of correlations between independent variables included in the model. The correlations between those variables should be below 0.7 [50]. Table 8 shows these correlations. The variables are shown on the horizontal and the vertical axis. The correlations between the variables are shown at the intersections. The number of rooms (5 classes) and the gross floor area (7 classes) have a high correlation which is explainable. The building type correlates quite high with the number of rooms, which makes sense if we think of detached houses having more rooms than flats. The same goes for building type and gross floor area, detached houses are in general bigger than terraced houses.

*Table 8. Correlations of assumed independent variables.*

| | Building type | Theoretical total energy use (kWh) | Number of rooms (5 classes) | Hot water heater type | Gross floor area (7 classes) | Garage or carport | Total actual energy use |
|---|---|---|---|---|---|---|---|
| **Building type** | | .380 | .698 | -.099 | .653 | -.114 | .520 |
| **Theoretical total energy use (kWh)** | .380 | | .347 | -.005 | .377 | -.048 | .414 |
| **Number of rooms (5 classes)** | .698 | .347 | | -.030 | .698 | -.077 | .490 |
| **Hot water heater type** | -.099 | -.005 | -.030 | | -.043 | -.026 | .100 |
| **Gross floor area (7 classes)** | .653 | .377 | .698 | -.043 | | -.202 | .461 |
| **Garage or carport** | -.114 | -.048 | -.077 | -.026 | -.202 | | -.113 |
| **Total actual energy use** | .520 | .414 | .490 | .100 | .461 | -.113 | |

The third assumption, the independence of observations is checked with the Durbin-Watson statistic and should be between 1.5 and 2.5 to assume there is no first order linear auto-correlation in the multiple linear regression data) [50]. The Durbin-Watson statistic is displayed in the rightmost column of  Table 4. The value is 1.956, which shows that it can be assumed that the third assumption is met.

Assumption 4, the linear relationship between the dependent variables and each of the independent variables is shown in the normal probability-probability plot, in Figure 4. The points should be approximately on the line, normality can be assumed since there are no drastic deviations [50].

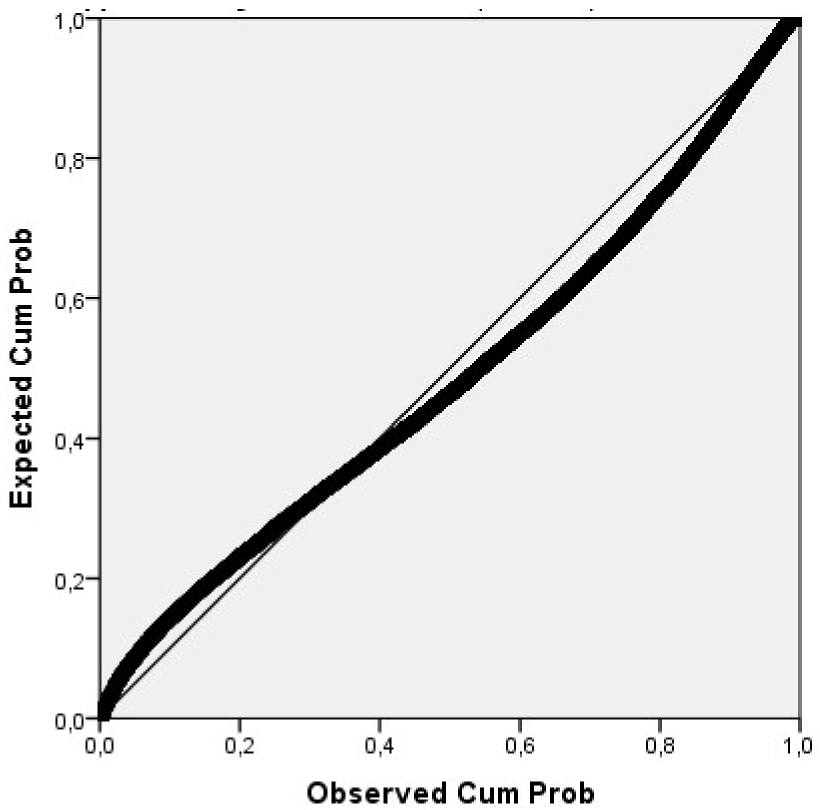


*Figure 4. The normal probability-probability plot of regression standardized residual. The dependent variable: total energy use.*

Assumption 5. To check for homoscedasticity the residuals scatterplot is presented in Figure 5. The data point should be scattered across the sample [50]. The data points are somewhat more scattered on the right of the plot but overall the distribution looks good enough to assume homoscedasticity.

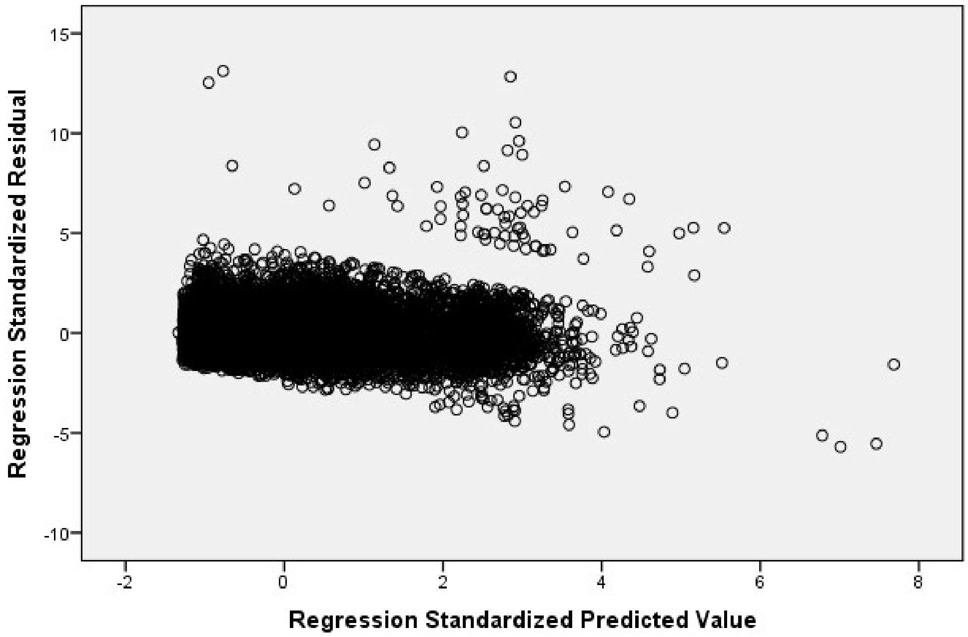


*Figure 5. Scatterplot of the residuals. The dependent variable: total actual energy use.*

Assumption 6. The absence of multicollinearity is confirmed by looking at the right end of the coefficients table in Appendix C. All the VIF values are below 10 indicating that this assumption is met [50].

The seventh assumption is met. To perform multiple linear regression, outliers in the sample need to be excluded. So, outliers are excluded using two criteria. Sample points with a z-score above 1.96 and below -1.96 are excluded from the sample. Also sample points with a Cook's distance greater than 1 are excluded from the sample to exclude multivariate outliers [50].

The final assumption to be tested is that residuals should be approximately normally distributed [50]. In Figure 6 the distribution of the residuals histogram is shown. The figure shows approximately normally distributed residuals and therefore this assumption is met.

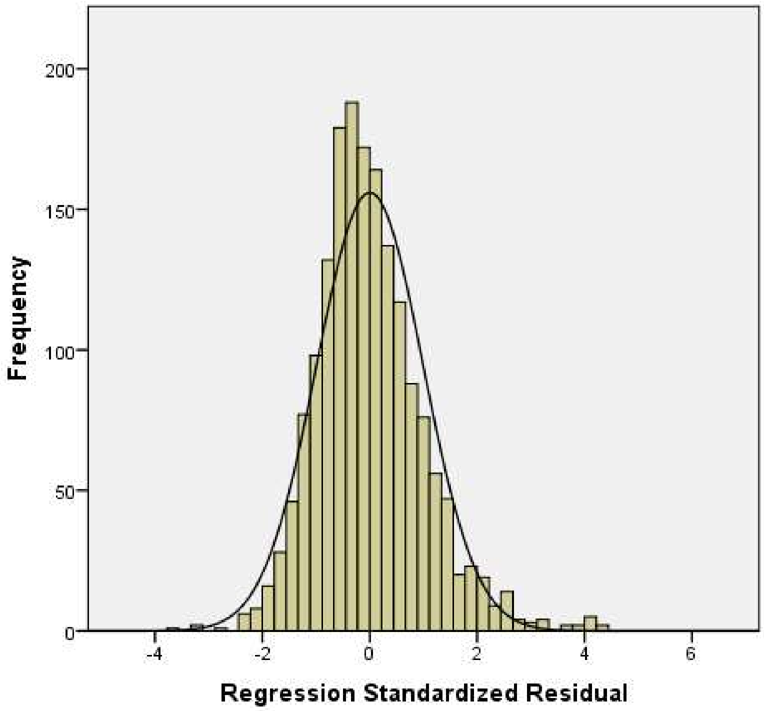


*Figure 6. Distribution of residuals histogram. Dependent variable: total energy use.*

In the coefficients Table 10 and Table 11 the sig. column all values are below 0.05 which shows the statistical significance of the models.

The 0 hypothesis is rejected since the significance in the ANOVA Table 9 under column 'Sig.' is below 0.05. Which means that all models are significant in predicting the actual energy use.

*Table 9. ANOVA of multiple regression analysis[3].*

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1,949E+10 | 6 | 3248635702 | 160,691 | ,000[b] |
| | Residual | 3,516E+10 | 1739 | 20216664,10 | | |
| | Total | 5,465E+10 | 1745 | | | |
| 2 | Regression | 2,221E+10 | 7 | 3173280338 | 170,034 | ,000[c] |
| | Residual | 3,244E+10 | 1738 | 18662618,36 | | |
| | Total | 5,465E+10 | 1745 | | | |
| 3 | Regression | 2,319E+10 | 11 | 2107835343 | 116,170 | ,000[d] |
| | Residual | 3,146E+10 | 1734 | 18144408,48 | | |
| | Total | 5,465E+10 | 1745 | | | |
| 4 | Regression | 2,394E+10 | 18 | 1329928762 | 74,790 | ,000[e] |
| | Residual | 3,071E+10 | 1727 | 17782209,24 | | |
| | Total | 5,465E+10 | 1745 | | | |
| 5 | Regression | 2,416E+10 | 24 | 1006548607 | 56,812 | ,000[f] |
| | Residual | 3,049E+10 | 1721 | 17717272,81 | | |
| | Total | 5,465E+10 | 1745 | | | |
| 6 | Regression | 2,431E+10 | 26 | 934998897,4 | 52,977 | ,000[g] |
| | Residual | 3,034E+10 | 1719 | 17648994,62 | | |
| | Total | 5,465E+10 | 1745 | | | |

[3] Dependent variable: total energy use. B: (constant) and building type), c: (constant, building type and theoretical energy use), d: (constant, building type, theoretical energy use and number of rooms), e: (constant, building type, theoretical energy use, number of rooms, and hot water heater type), f: (constant, building type, theoretical energy use, number of rooms, hot water heater type, and gross floor area), g: (constant, building type, theoretical energy use, number of rooms, hot water heater type, gross floor area, and garage or carport).

## Appendix C

Coefficients table of the multiple linear regression models 3 and 6. This section presents the coefficients tables with the unstandardized and standardized coefficients, the significance level of the variables (Sig.), the lower and upper boundaries of the 95% confidence intervals and the collinearity statistics.

*Table 10. Coefficients, significance, confidence interval and the collinearity statistics of linear regression model 3.*

| Approximately 80% of the cases (SAMPLE) | Unstandardized Coefficients | | Standardized Coefficients | | 95,0% Confidence Interval for B | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | B | Std. Error | Beta | Sig. | Lower Bound | Upper Bound | Tolerance | VIF |
| (constant) | 8809,988 | 129,358 | | 0,000 | 8556,435 | 9063,541 | | |
| terraced house | 1945,417 | 109,882 | ,138 | ,000 | 1730,039 | 2160,795 | ,560 | 1,787 |
| Semi-detached house | 4079,548 | 189,047 | ,147 | ,000 | 3708,999 | 4450,096 | ,742 | 1,348 |
| detached house | 9418,333 | 188,787 | ,362 | 0,000 | 9048,294 | 9788,372 | ,650 | 1,540 |
| farm | 2875,069 | 372,507 | ,048 | ,000 | 2144,923 | 3605,215 | ,867 | 1,153 |
| Home with store | 2598,936 | 476,235 | ,033 | ,000 | 1665,474 | 3532,399 | ,934 | 1,071 |
| residential building with communal facilities | 3449,658 | 799,102 | ,026 | ,000 | 1883,348 | 5015,969 | ,958 | 1,044 |
| theoretical total energy use | ,428 | ,014 | ,207 | ,000 | ,400 | ,456 | ,736 | 1,358 |
| rooms (3) | -173,859 | 143,519 | -,011 | ,226 | -455,169 | 107,452 | ,406 | 2,464 |
| rooms (4) | 1053,856 | 154,294 | ,071 | ,000 | 751,426 | 1356,286 | ,315 | 3,179 |
| rooms (5) | 2031,436 | 175,625 | ,115 | ,000 | 1687,195 | 2375,677 | ,348 | 2,870 |
| rooms (6+) | 3543,129 | 204,118 | ,158 | ,000 | 3143,040 | 3943,219 | ,414 | 2,413 |

*Table 11. Coefficients, significance, confidence interval and the collinearity statistics of linear regression model 6.*

| Approximately 80% of the cases (SAMPLE) | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. | 95,0% Confidence Interval for B Lower Bound | 95,0% Confidence Interval for B Upper Bound | Collinearity Statistics Tolerance | Collinearity Statistics VIF |
|---|---|---|---|---|---|---|---|---|---|
| (constant) | 8875,328 | 176,736 | | 50,218 | 0,000 | 8528,910 | 9221,747 | | |
| terraced house | 1983,857 | 118,944 | ,141 | 16,679 | ,000 | 1750,717 | 2216,997 | ,474 | 2,111 |
| semidetached house | 3578,691 | 207,470 | ,129 | 17,249 | ,000 | 3172,031 | 3985,350 | ,611 | 1,637 |
| detached house | 8512,163 | 218,958 | ,327 | 38,876 | 0,000 | 8082,985 | 8941,341 | ,479 | 2,088 |
| farm | 1694,939 | 393,877 | ,029 | 4,303 | ,000 | 922,906 | 2466,973 | ,770 | 1,299 |
| home with store | 1772,249 | 484,922 | ,023 | 3,655 | ,000 | 821,758 | 2722,740 | ,893 | 1,119 |
| residential building with communal facilities | 3976,247 | 807,674 | ,030 | 4,923 | ,000 | 2393,134 | 5559,360 | ,930 | 1,076 |
| theoretical total energy use | ,377 | ,015 | ,182 | 25,332 | ,000 | ,348 | ,406 | ,659 | 1,518 |
| rooms (3) | -173,968 | 157,923 | -,011 | -1,102 | ,271 | -483,511 | 135,574 | ,332 | 3,008 |
| rooms (4) | 902,836 | 173,033 | ,061 | 5,218 | ,000 | 563,676 | 1241,996 | ,248 | 4,031 |
| rooms (5) | 1678,422 | 194,662 | ,095 | 8,622 | ,000 | 1296,868 | 2059,976 | ,281 | 3,555 |
| rooms (6+) | 2952,529 | 221,814 | ,131 | 13,311 | ,000 | 2517,753 | 3387,305 | ,348 | 2,873 |
| CV and solar water heater | -337,272 | 316,716 | -,006 | -1,065 | ,287 | -958,063 | 283,518 | ,993 | 1,007 |
| Gas water heater | 871,756 | 246,569 | ,021 | 3,536 | ,000 | 388,459 | 1355,052 | ,978 | 1,023 |
| electrical boiler | 669,680 | 156,813 | ,025 | 4,271 | ,000 | 362,313 | 977,048 | ,964 | 1,038 |
| heat pump water heater | -169,829 | 346,658 | -,003 | -,490 | ,624 | -849,309 | 509,651 | ,993 | 1,007 |
| block or district water heater | 1317,854 | 235,873 | ,033 | 5,587 | ,000 | 855,522 | 1780,186 | ,947 | 1,056 |
| district water heater | -63,847 | 191,511 | -,002 | -,333 | ,739 | -439,226 | 311,532 | ,977 | 1,024 |
| no water heater | 190,354 | 2487,333 | ,000 | ,077 | ,939 | -4685,040 | 5065,749 | ,997 | 1,003 |
| GFA (50-69) | -313,462 | 192,398 | -,015 | -1,629 | ,103 | -690,579 | 63,655 | ,385 | 2,597 |
| GFA (70-89) | -107,912 | 191,250 | -,007 | -,564 | ,573 | -482,778 | 266,954 | ,247 | 4,041 |
| GFA (90-119) | 264,665 | 199,733 | ,018 | 1,325 | ,185 | -126,829 | 656,159 | ,186 | 5,374 |
| GFA (120-149) | 669,580 | 228,825 | ,032 | 2,926 | ,003 | 221,062 | 1118,097 | ,276 | 3,625 |
| GFA (150-199) | 1176,111 | 269,559 | ,043 | 4,363 | ,000 | 647,751 | 1704,471 | ,355 | 2,814 |
| GFA (200+) | 2208,334 | 321,469 | ,064 | 6,870 | ,000 | 1578,226 | 2838,441 | ,392 | 2,549 |
| garage | 487,761 | 132,165 | ,028 | 3,691 | ,000 | 228,705 | 746,816 | ,596 | 1,677 |
| carport | 867,893 | 242,258 | ,021 | 3,583 | ,000 | 393,045 | 1342,740 | ,951 | 1,052 |