Utrecht University

Towards a Data-Driven Energy Advice

An Exploratory Analysis of Gas Usage and Resident's Behavior

Lesley Joy Rietvelt
4352327
Master Thesis
MSc Applied Data Science
Utrecht University
First reader: Prof. dr. Arno Siebes
Second reader: dr. Ing. habil G. Krempl
2 July 2021

Word count: 8,632

Table of Contents

Preface

This study was carried out as an assignment for the company Intergas and is divided into two parts. The first part is a collaboration of three MSc Applied Data Science students from University Utrecht. This part aims to create a base prediction model that predicts gas usage using the available data. Within this collaboration part, we have made use of each student's own strengths. This resulted in the following division of tasks:

Name	**Responsibility**
Martin Doornhein	Data analysis in Python, development of methodology of regressor model implementation.
Robin Reijers	Background research with a focus on the energy gap and exploratory data analysis.
Lesley Rietvelt	Background research with a focus on the current energy label and previous research.

Tasks such as writing, creating figures, and tables were often done in collaboration. Next to this, all decisions concerning the base model and writing were discussed among each other. The second part of this thesis consists of the individual research. In this case, the research on detecting absence is presented.

Introduction

The European Union has set the goal to become climate neutral by 2050. Each country that falls under the EU, is required to develop national policies that support this goal (RvO, n.d.). In the Netherlands, the The Netherlands Enterprise Agency (RvO) is responsible for implementing these national policies. According to the EU, homes and utility buildings are responsible for 40 percent of our energy consumption and 36 percent of total CO_2 emissions (European Commission, 2020). Therefore, the RvO is implementing policies that help lower the CO_2 emissions of buildings and improve their energy efficiency.

One of these policies is the energy labelling system, which is currently used to advise home owners on their house's energy efficiency (RvO, 2021). To stimulate homeowners to invest in improving energy efficiency and to make this process accessible, the RvO is trying to create a National Digital Platform. Their first step is to automate the process of determining the energy efficiency of Dutch houses. This means the current energy labelling system will be converted into a digital energy label.

In the Netherlands, homeowners are required to provide an energy label upon delivery, sale or rental of the home. When a homeowner wants to register the energy label, an energy advisor inspects their home and calculates its energy efficiency. Eventually, the energy label indicates the overall performance of the house from A++++ to G. Next to this, it contains details on the insulation and installations, offers advice for future improvement, and estimates the energy costs of the homeowner (RvO, 2021).[1]

On 1 January 2021, the method that determines the energy efficiency of a house was updated. Energy labels are now held against new standards and documents (NTA 8800, BRL 9500, and ISSO 82.1) (RvO 2021). Although the new and old methods overlap, NTA 8800 requires more information on the building and advisors have to report every apartment separately. Next to this, more detailed descriptions of, among others, pipe transits and insulation, heating installations, delivery and distribution systems, and cooling installations are required (RvO, 2020a). For example, DGMR[2] established that the construction method of a building influences its energy efficiency, and therefore advisors are required to provide information on the construction method from January 2021 onwards (RvO, 2020b). Thus, creating an energy label with the newly required standards and documents is a time-consuming process. An inspection takes, on average, one to two hours to record all the details mentioned above (Energielabel, n.d.). This extensive inspection should theoretically guarantee accuracy. However, the label given is highly influenced by the methods and precision of the advisor and the availability of certain information (for example, whether or not it is possible to retrieve what

[1] For an example of a 2021 energy label (in Dutch), see: https://www.RvO.nl/sites/default/files/2020/12/energielabel-voorbeeld-woningen.pdf.

[2] DGMR is a Dutch, independent consultancy firm on building, industry, traffic, environment, and software. See: https://dgmr.nl/en/.

materials were used for improving insulation) (Radar, 2021). Next to this, energy labels do not accurately reflect the energy consumption of a house due to the "Energy Gap".

Recently, an increasing amount of research has been done regarding energy labels and the "Energy Gap". This research shows energy labels give an estimate of the energy consumption that often differs significantly from the actual consumption. This difference is called the energy gap. Buildings that are considered inefficient have a lower actual consumption than the estimated consumption, while buildings that are considered energy efficient often have a higher actual consumption than the estimation (Majcen, Itard & Visscher, 2012; van den Brom, Meijer & Visscher, 2018). Policies to reduce energy consumption in order to meet the 2050 goal are based on the theoretical consumption and energy efficiency of buildings. Since these estimations differ from the actual energy efficiency, these policies are less effective in practice than on paper (Macjen, Itard & Visscher, 2012). Several studies called for a better way to address the current energy gap and the usage of energy labels in housing as houses are currently labeled unfairly (Majcen, Itard & Visscher, 2013; van den Brom, Meijer & Visscher, 2018; Boonekamp, 2007; Martens & Spaargaren, 2005; Majcen, Itard & Visscher, 2015).

Intergas is researching the possibility of creating a digital energy labelling method based on their data, which could perhaps decrease the energy gap. Intergas is a company that focuses on developing sustainable heating systems, advising customers based on data, and selling their energy efficient systems (Intergas, n.d.). Over the years, they have collected data from their heaters, resulting in approximately 6TB of information in their database. This data, together with weather measurements from the Royal Dutch Meteorological Institute (KNMI), can be used to create basic and improved models that are able to classify houses based on their energy usage. Since such a model would be based on actual gas consumption, it is possible the estimations are more accurate than current methods.

Earlier research shows several factors influence actual gas usage and energy efficiency. For example, building characteristics such as the surface of the building's floors in m^2 and building year often have an impact on energy consumption (Majcen & Itard, 2014a). Resident's characteristics (such as age, income and social standing) also influence the consumption, which was apparent due to differences in consumption between houses with similar characteristics (Majcen, Itard & Visscher, 2013; Yun & Steemers, 2011; Berkland, 2014; Santin, 2011; Jeeninga, Uyterlinde & Uitzinger, 2001). Lastly, resident's behavior (such as heating patterns, airing and absence) has an impact on how much energy is used as well (Majcen & Itard, 2014b). The data provided by Intergas contains information on some of the important building characteristics, such as the building year and surface area. Next to this, Intergas' data could be used to gain insight into resident's behavior by trying to detect heating patterns and research possibilities for correcting for this behavior.

Research also shows electrical usage is negligible when researching energy labels. The RvO energy label is exclusively taking electrical installations into consideration, while most

electricity usage comes from household appliances. The theoretical electricity usage and actual electricity usage of those installations do not differ from each other, or per label class (Majcen & Itard, 2014b). Therefore, electricity usage is negligible in this analysis.

This thesis then aims to explore the opportunities of creating a model based on Intergas' data which can accurately predict the gas consumption of a building in order to eventually assign correct energy labels. Next to this, the individual cases will research possibilities of clustering, classifying and correcting in order to improve the base model. If such a model can indeed be created, future research can improve the accuracy of this model and further explore the possibilities of creating a digital energy label, but that is not within the scope of this project.

Next to a base model, this thesis aims to detect resident's absence by analyzing the resident's heating behavior. As mentioned before, resident's behavior influences gas consumption and therefore it is interesting to explore whether or not it is possible to gain insight on resident behavior and add information to the existing database that could benefit the base model. For example, within this research it is attempted to create conditions to classify rows (representing daily averages and daily gas usage) as "absent" or "present". A classification column could be added to the existing data frame for future reference.

Base Model: Methodology

Since the aim of this project is exploratory, the focus laid on what information the data could provide. However, the large amount of data forced us to make decisions about filtering and rearranging the data. Next, exploratory analysis was done to find correlations and extract information. Eventually, a basic model was created and separate analyses were done to explore further possibilities. The software, data and methods used during this project, together with the decisions that have been made, will be explained in more detail in this chapter.

Software

Intergas' database was built upon Hadoop and consists of approximately 6TB of data. With large amounts of data, it is recommended to use Apache Spark. Spark is fast, offers less reading and writing from and to the disk and, due to the Python API PySpark, is fairly easy to use (Honold, 2020). It was also possible to work on smaller portions of the data, meaning other Python packages (e.g. Pandas) could be used. Overall, the following software packages were used: PySpark, Pandas, NumPy, Matplotlib, Seaborn, and Plotnine.

Data

The data provided was spread among different datasets. These datasets were used to create a data frame that would be useful for different analyses (see figure 1). This paragraph will discuss the different datasets and the features in it.

Intergas_raw

This dataset contained many columns. Most of its content was about small interval measurement values. For this project, the focus was on 24 hour data. Therefore, a selection of only a few columns was made, which gave information on the gas usage of central heating and warm water. Next to this, all gas use measurements below zero were removed when loading this dataset. This was done to reduce the amount of errors in the data and decrease the computational load while merging datasets.

Gas_use_hourly

This dataset contained information about hourly gas used per m^2 grouped by heater id's. These id's are unique numbers per heater (see table 3).

KNMI_data_24

This dataset is provided by Intergas. This dataset contained information about the weather for every hour, grouped by time point and neighborhood.

Ig_heater_info

This dataset contained heater ID's, the neighborhood and two house properties from the given house. These properties were house building year and total surface area of residence.

First, the datasets described above were merged to create a data frame that could be used for inspection, exploration and analysis. This was done by using inner join, meaning the intersection of both datasets is used for the newly created data frame. Since this method only adds the overlapping keys, the new data frame has a low amount of missing values. Reducing the number of missing values was desirable, due to the small number of columns. Table 1 shows the order that was used to create the data frame for further analysis.

Table 1. Joining steps for the final dataset

Dataframe 1	**Dataframe 2**	**Keys**	**New dataframe**
Intergas Raw	Ig_heater_info	Heater_id	Merge_1
Merge_1	Gas_use_hourly	Heater_id, date_day	Merge_2
Merge_2	df_knmi	Neighborhood, date_day	df_24hour

This final data frame contained 28,651,724 rows and 12 columns (see table 2 for information on the variables). From this data frame, a sample of 10 percent was taken. Since the whole data frame is too large to work with easily, this sample will be used for testing analysis and visual inspection.

A final step that was performed before the filtering process, was dealing with missing values. Based on the first inspection, it seemed that there were missings in four columns (see table 3). These columns were: rain, sun, temp and building year. After closer inspection, it became clear that there was a fifth column that had missings which were valued with a zero (n = 74,114). This column was Surface and the missings did overlap 100 percent with the building year column.

Table 2. Variables in final data set (n = 26,393,420)

Variable_name	Type	Meaning
Heater_id	Integer	Unique heater identification number
Neighborhood	Integer	Neighborhood
Date_day	Integer	Year/month/day
Surface	Integer	Surface of total house in m^2
Building_year	Integer	Building year of house
Rain	Double	Amount of rain in 0.1 mm (precipitation <0.05mm = -1)
Sun	Double	Amount of sun in 0.1 hours (sunshine <0.05h = -1)
Temp	Double	Temperature in celsius * 10
Wind	Double	Wind in 0.1 meters / second
T_act	Double	Actual inside house temperature (°C)
T_set	Double	Set inside house temperature (°C)
Gas_ch	Double	Gas use in m^3 for heating house
Gas_dhw	Double	Gas use in m^3 for hot water
Temp_diff	Double	Inside temperature - outside temperature (°C)

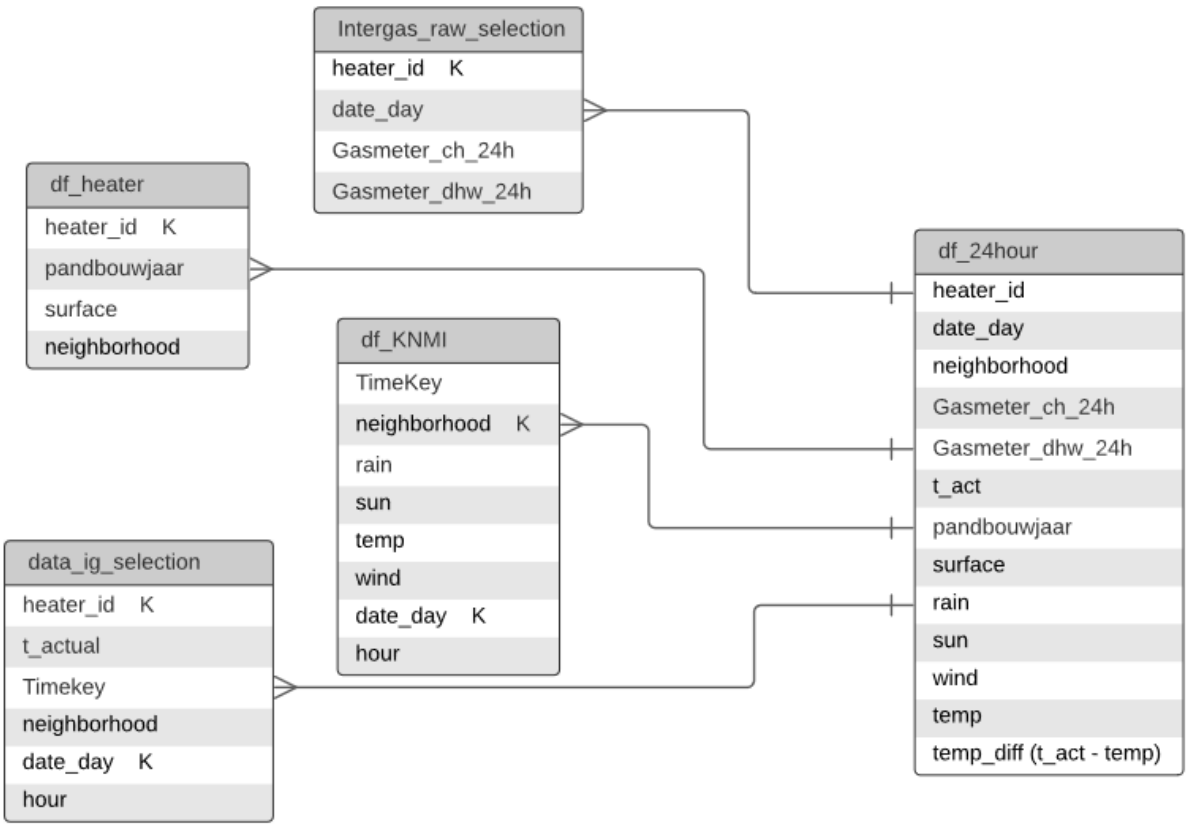
Table 3. Columns with missing values

rain	sun	temp	building year
91,229	279,904	44,881	74,114

The missings values came from two different datasets that were used in the merging of datasets. The missings in the weather variables came from the KNMI dataset. A possible explanation for this missings is an error with one, or more of the weather stations. The missings in the variable building year and surface came from the Ig_heater_info dataset. A reasonable explanation for these missings is that these customers did not want to communicate their information. Due to not having a proper NMAR analysis, it was not possible to reverse calculate the missing data. To prevent skewing of the data it was chosen to leave out the rows which contained missing data.

After removing missing values (n = 259,922), the number of rows was: 28,393,420. The missing values were divided among rain (86,290), sun (258,923), temperature (42,533). For building year and surface 74,114 values are missing.

Figure 1. Data frames that were used to get the final dataframe



Filters

Table 4 shows the filters that were used to clean the dataset. These filters were chosen based on a combination of reasoning and the influence of the filter on the data. For the variable surface, the cutoff value of 40 was used. Houses smaller than 40 square meters were removed because these values are very unlikely to be correct. Looking at the data, 40 seemed to be an appropriate cutoff because the number of removals highly increased after applying a value higher than 40.

For both gas_measures, a cutoff value of 40 m^3 per 24 hours was used. Based on information from Intergas, values higher than 40 m^3 are very unlikely and could be considered errors from the heater. These errors can be caused by a momentary loss of power, for example. From the data perspective, these cutoffs did not influence the data much.

For the variables t_act (temperature measured inside) and t_set (temperature set on the thermometer), the cutoff value is equal or smaller than 26 degrees Celsius. These points were chosen because values higher than 26 are unusual. For t_act, there was also a minimum filter with a value of 10 added. This cutoff value was selected based on the data.

Table 4. Filters and number of removed rows after applying

Variable name	Filter	# removed	% removed
Surface min	> 40	299464	1.05%
Surface max	< 600	269998	0.94%
Gas_ch	< 40	7286	0.03%
Gas_dhw	< 40	484	0.002%
T_act	<= 30	33607	0.12
T_act	>= 10	1393515	4.86%
T_set	<= 26	19060	0.07%

The total number of removed rows is 1,978,666 which is 6.91% of the total number of rows (see table 5 for the descriptives). Lastly, the variable temp_diff (temperature difference) contained values below zero, meaning the outside temperature was higher at these moments than the inside temperature. In this case, there is no gas needed. Therefore, the limit of this variable was modified and all negative values were replaced with zero.

Table 5. Descriptives of the removed rows

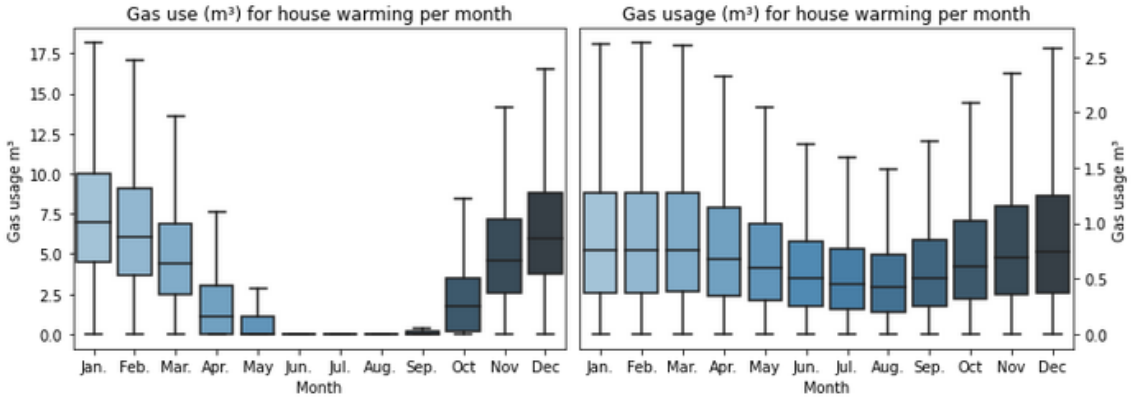
Summary	Surface	Gas_ch	Gas_dhw	T_act	T_set
Count	1.978.666	1.978.666	1.978.666	1.978.666	1.978.666
Mean	388,27	23,99	34,89	6,56	5,47
Standard deviation	1.553,91	2.471,23	3.610,76	11,27	8,77
Minimum	0	0	0	-323,65	-154,27
Maximum	68.353,0	429.496,73	429.483,71	326,95	325,11

Exploratory data analysis

After applying the filters, the dataset was reduced by a total of 6.91 percent when compared to the total data available. With this exploratory data analysis the make-up of the dataset will be shown. All visualisations are based on the earlier mentioned sample of 10 percent.

Gas usage for warming houses is lower (0.18 ± 0.73) in the summer months and higher in the winter months (7.10 ± 4.56). This is also true for warm water (0.57 ± 0.50) in summer versus (0.91 ± 0.78) in winter) but with not as much of a difference (figure 2).

Figure 2: Boxplots of gas usage per 24 hours for house warming (left) and warm water (right)



During the exploratory analysis, correlations were found between certain variables. For example, there is a positive correlation between the average sunshine per day and average temperature (0.68). Another positive correlation (0.53) is temperature difference inside/outside and gas_ch (gas used for heating purposes).

With negative correlations the avg(temp) and temp(diff) are strongly negatively correlated (-0.93), also temperature difference inside/outside and average sunshine are negatively correlated (-0.60). Other negative correlations are between gas usage for heating and average temperature (-0.58) and the average sunshine (-0.47).

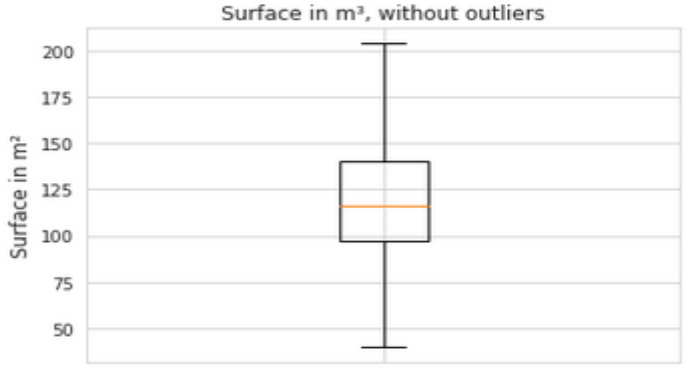
Figure 3: Correlations between variables

	gas_ch	gas_dhw	building_year	surface	rain	sun	temp	wind	t_act	temp_diff
gas_ch	1,00	0,18	-0,05	0,13	0,02	-0,47	-0,58	0,13	-0,33	0,53
gas_dhw	0,18	1,00	0,03	0,05	0,00	-0,12	-0,16	0,06	-0,02	0,17
building_year	-0,05	0,03	1,00	0,03	0,00	0,00	0,01	0,02	0,06	0,02
surface	0,13	0,05	0,03	1,00	0,00	0,02	0,03	-0,02	-0,03	-0,05
rain	0,02	0,00	0,00	0,00	1,00	-0,18	0,07	0,23	-0,02	-0,09
sun	-0,47	-0,12	0,00	0,02	-0,18	1,00	0,68	-0,29	0,42	-0,60
temp	-0,58	-0,16	0,01	0,03	0,07	0,68	1,00	-0,19	0,51	-0,93
wind	0,13	0,06	0,02	-0,02	0,23	-0,29	-0,19	1,00	-0,13	0,16
t_act	-0,33	-0,02	0,06	-0,03	-0,02	0,42	0,51	-0,13	1,00	-0,15
temp_diff	0,53	0,17	0,02	-0,05	-0,09	-0,60	-0,93	0,16	-0,15	1,00

The most frequently occurring surface area is 110m² with the lowest value being 40 m² due to the filter applied on surface area and the highest value is 600m². The average surface area in square metres is 125.42 (± 52). The outliers are not shown in the boxplot to reduce visual clutter (Figure 4).

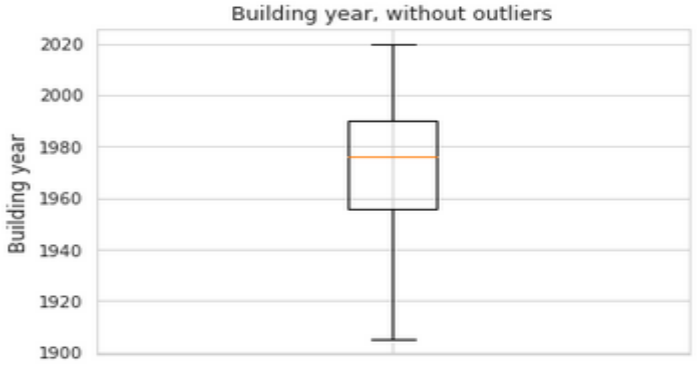
Most of the data from Intergas is from residential housing, this would explain the distribution of surface area in housing that is found in the dataset. Meaning that the model possibly could face difficulties when trying to estimate the actual gas consumption in houses with a larger surface area.

Figure 4: Boxplot of housing per m²



In the dataset the most frequently occurring building year of houses is 1978. The oldest building in the dataset dates from the year 1300 and there are no buildings newer than being built in 2020. Building years of the houses are on average 1971 (± 62) and there are no outliers above 2020 in the dataset (figure 5). This range of years encompasses several different updates of the EU regulations in housing insulation and requirements. Meaning that within the dataset there are houses which adhere to different standards of insulation and specifications.

Figure 5: Boxplot of building years



Within the data it is found that summer months (Jun - Aug) have the highest temperatures and the most amount of sun hours for each month. While the winter months (Dec - Feb) have the lowest temperatures and amount of sunshine each month (figure 6 and 7). This tells us the data

itself is in line with the weather patterns as reported from the KNMI and that there are no abnormalities within the general sense of the dataset.

Figure 6: Boxplot of average temperature

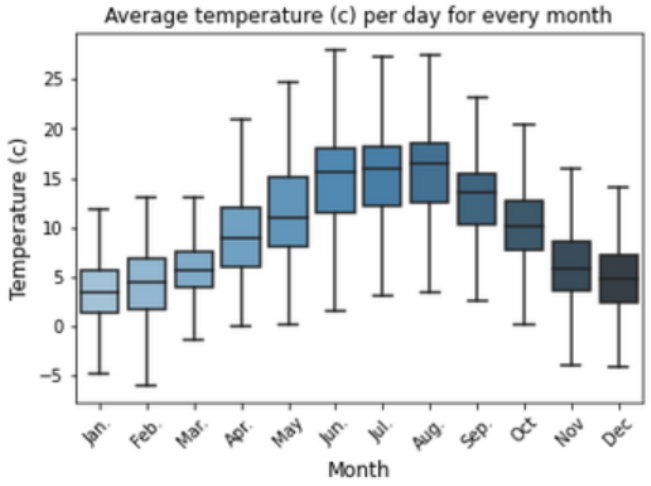
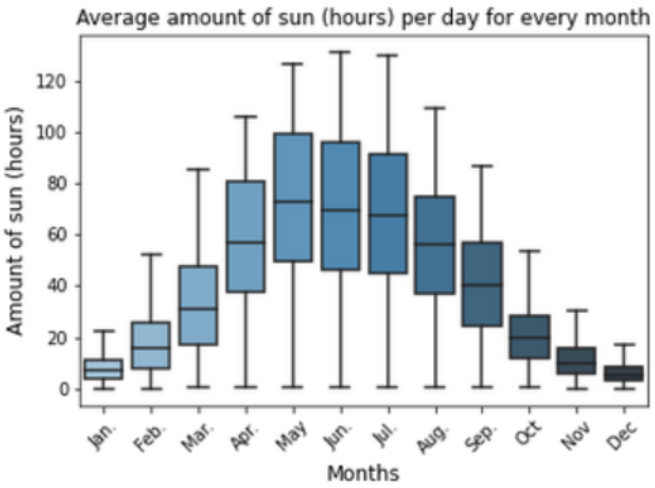
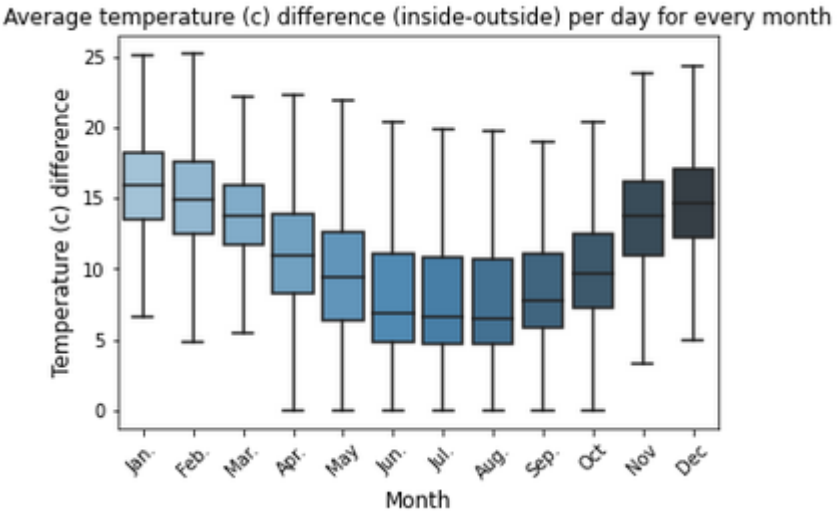


Figure 7: Boxplot of average sun hours



For the average temperature difference (inside - outside) the greatest temperature difference is found in the winter months and when looking at the smallest temperature difference it is found during the summer months (figure 8).

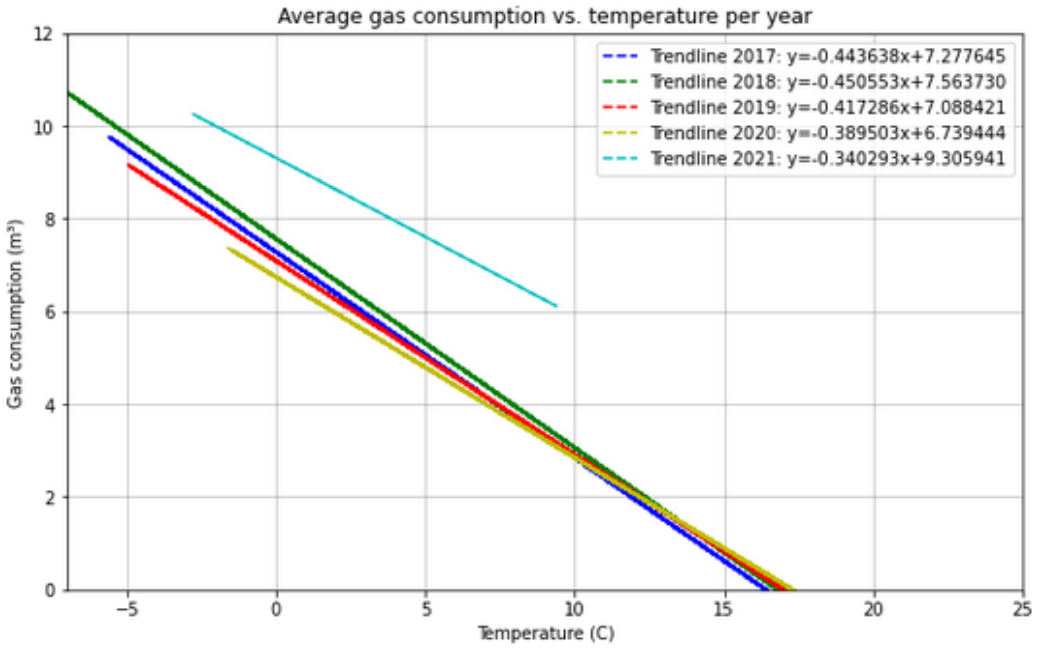
Figure 8: Average temperature difference (inside - outside) in Celsius per day for every month



For all years there is a negative correlation found between gas consumption and temperature outside. The trendlines are linear, which is expected due to specific heat. Specific heat "is defined as the energy required to raise the temperature of a unit mass by one degree" (Yunus, Cengel & Ghajar, 2020). It costs the same amount of energy to heat up the air in a room from 0 to 1 degree Celsius as it does from 5 to 6 degrees. On average the years 2017 through 2020 each

have a similar slope and starting point, but the year of 2021 has a way higher starting point and slope value. This is due the dataset only partially containing 2021 and only the winter months, in these months the gas consumption is the most as previously shown (figure 9). Next to this, we expect the average gas usage to be higher due to working from home during corona.

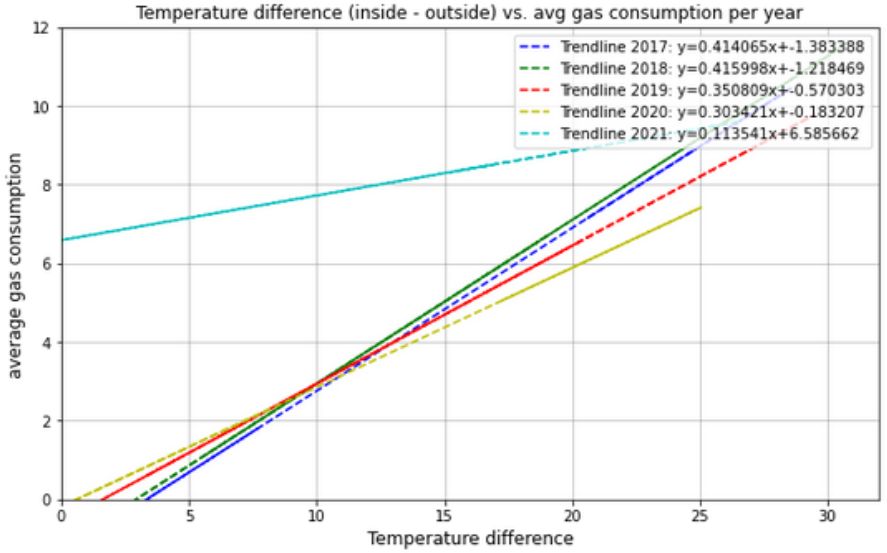
Figure 9: Average gas consumption plotted against temperature per year



The average gas consumption per year is positively correlated with the total temperature difference between inside and outside of homes. Meaning that the higher the temperature difference is, the higher the gas consumption is, tying together the temperatures outside during colder periods with the increased gas consumption to gain the same amount of heating as one normally would get with higher outside temperatures.

All the years between 2017 and 2020 show a similar slope profile. In 2021 there is a flatter but increased slope value due to only partly having values from the winter months and not the rest of the year meaning that the data is skewed to more gas consumption then in the previous years (figure 10).

Figure 10: Average gas consumption per year plotted against temperature difference (inside – outside)



Modelling base model

To model the data and predict the gas usage, a random forest regressor model was used. This model aims to predict the total gas usage per year, per m^2. The Random Forest regressor was chosen for different reasons. First, the model copes well with non-linear relations. Second, it deals well with collinearity. This is desirable because the KNMI-predictors did correlate among each other, for example: sun and temp (r = .701). One of the drawbacks of this method is that it is not able to extrapolate predictions. A linear regression model could predict outside the range of the train set.

To feed this random forest model, the earlier described dataframe was grouped by heater_id and years. This resulted in a dataframe with averages of all the variables for every heater id, per year. The outcome variable total gas consumption per m^2 was included. Some columns, which were no longer of need, were removed (heater_id, year, gas_ch, gas_dhw). To get the data in the right format to train the regressor, the vectorindexer from the PySpark package was used.

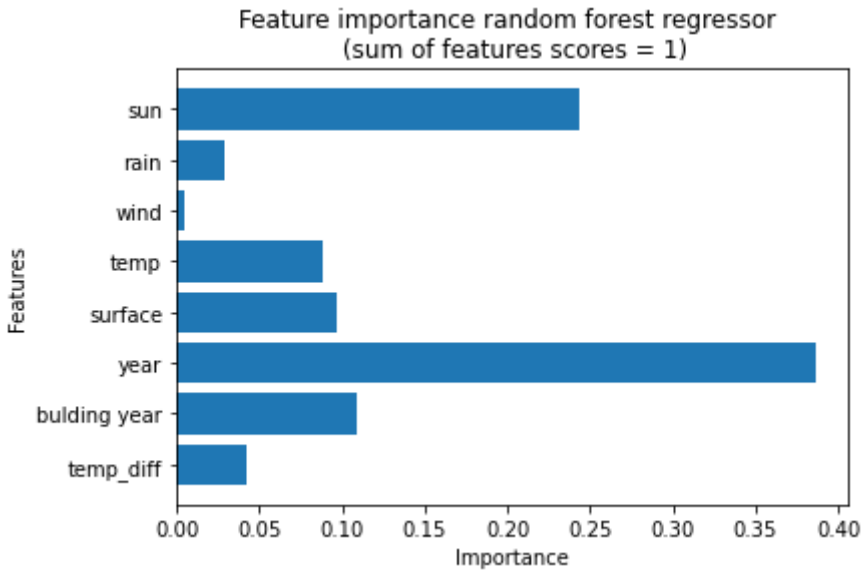
To train and test the model, a proportion of the data (30%) was held out for validation. The train set contained 87,399 rows and the test set contained 37,388 rows. Because there was a lot of data available, the hold out method seemed to be sufficient as a method to determine testing accuracy. Cross validation would be an alternative, but because of the high amount of data, this method would be highly computational expensive.

Base Model: Results

Different random forest regressor models were performed. Every model is trained on the whole dataset. The initial random forest model (r^2 = .460, RMSE = 0.0230) has the goal of being a baseline model. This model aims to predict gas usage per m^2 and was trained on 124,787 observations, which was divided into a train set of 70 % (n = 87,399) and a test set of 30% (n = 37,388). The mean of the outcome column is: 0.044 . In an effort to increase accuracy, an extra filter was added. The dataset was filtered for the number of observations within a year. This filter is applied for the cutoff values > 200 and > 250. For these models, the same train-test-set proportions were applied. The first model (n = 70,993, r^2 = .239, RMSE = 0.015) and second model (n = 64,413, r^2 = .245, RMSE = 0.015) both performed less well and had a lower explained variance. Therefore, the first model is chosen to be the base model.

To inspect the impact of the features on the prediction accuracy, a feature importance plot was visualized (figure 11). The highest scoring features are year and sun. Two features had a relative high influence on the model which were: year and sun. The feature wind had the lowest impact on the model.

Figure 11: Feature importance barplot for random regressor model



Detecting Resident's Absence: Methodology

The base model created in the previous chapters can be considered a first step towards a digital energy label. From previous research it became clear gas usage is influenced by factors such as building year, floor surface, and resident's behavior (Macjen & Itard, 2014b). The first two factors can be taken into account when creating the base model. However, there is no data readily available on resident's behavior. Therefore, the next step in this research is to gain insight into resident's behaviors and explore whether or not it is possible to implement corrections for this behavior to improve the model's performance.

In the second part of this thesis, the possibilities of detecting and correcting for resident's absence will be explored as one of the many factors influencing resident's behavior. First, an exploratory analysis was done to gain insight on stoking behavior and create possible classification conditions. The data frame used for this analysis, its methodology and the preliminary results will be discussed in this chapter.

Data frame

The 24 hour data frame, as described in the earlier methodology chapter, was used. First, the months December, January and February were filtered to represent the winter. Since it is cold during these months, most residents use gas during the day. Therefore, being absent and using minimal gas would stand out and would be easier to recognize during analysis. The gas usage for the central heating would be most reliable during this period, since there are larger differences in usage. Finally, since the aim is to recognize absence, it is interesting to keep in mind the winter break and holidays (Christmas and New Year's) when it is more likely residents are absent for a few consecutive days. However, due to filtering, there are relatively little observations left for each month. For example, a heater had a maximum of 11 observations for January 2020, and a maximum of 13 observations for December 2019, meaning it is only possible to analyse the absence within these 11 to 13 days. This means it is not likely to find consecutive days of absence, which could have been classified as vacations, in the current data frame.

Exploring heater ID's

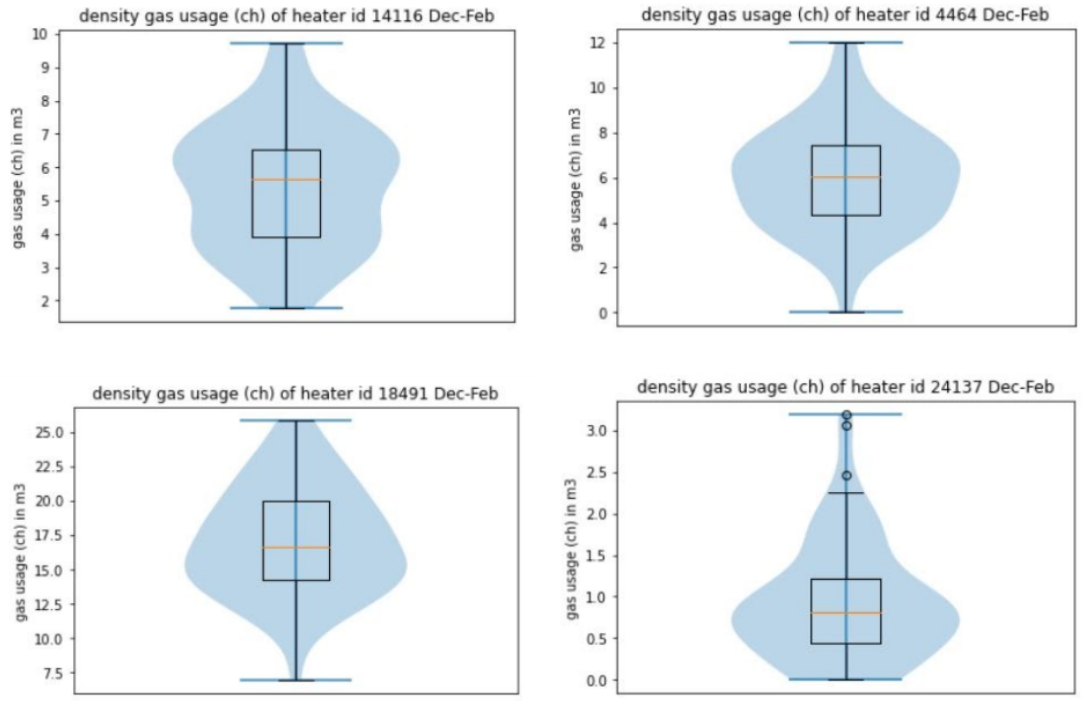
The first step of the exploratory analysis was to select a few heater ID's. In order to gain an understanding of their behavior, their usage was analyzed and documented (see table 6). Next, four heaters were selected based on the information in table 6, so the behavior of these residents could be compared to each other. These were heater_id's 14116 (due to its average usage and zero days without usage), 4464 (due to its average usage and one day without usage), 18491

(due to its high mean usage) and 24137 (due to its low usage and five days without central heating). These heater behaviors were plotted for comparison (see figure 12).

Table 6: Selected heater_id's from December to February based on number of entries

heater_id	# entries	mean gas usage (ch)	# entries below mean	# days without gas usage (ch)	# days without gas usage (dhw)
14116	75	5,43	35,00	0	0
4464	73	6,02	36,00	1	1
18491	72	17,076	38,00	0	0
14963	70	6,402	36,00	1	1
10021	70	5,617	35,00	0	0
18763	69	4,595	39,00	1	1
24137	69	0,935	42,00	5	0
11828	69	5,403	39,00	0	0
8744	68	7,774	37,00	0	0
2847	68	6,649	37,00	1	1

Figure 12: Density of the gas usage for central heating of December, January and February from four selected heaters



From these plots it is clear every heater_id, and thus every resident, behaves in its own way. It is therefore challenging to create conditions that are applicable for each heater. The next step was to extract more information from the heaters to see whether zero gas usage actually corresponds with absence. For this exploratory analysis other variables, such as t_set (temperature set inside), t_act (temperature measured inside) and temp_diff (temperature difference) were considered.

First, heater 24137 was taken for further inspection. From the information from table 7, it is possible to conclude there were at least five days with no gas usage, since there were five instances of zero gas usage for central heating. Although there was gas used for domestic hot water, this could still indicate absence of the resident. Many heaters nowadays have a "comfort mode" where the heater ensures quick access to hot water by always heating a small amount of water. To be able to keep this water warm gas is used, which explains the gas_usage (dhw) is almost never zero, but close to zero at times (Feenstra, 2019).

In order to establish whether or not the resident was absent during the 26th, 27th and 28th of February 2019, t_set and t_actual were analyzed. Since both these variables had higher values than expected (t_set = 18.9, 19.0, 17.5 and t_act = 20.7, 22.1, 21.8) it was concluded there was no absence. In the case of heater 24137 it is more likely there is an external heating source. From this short inspection, it is concluded that variables as t_set and t_act are important to consider when establishing whether or not a low gas usage corresponds with absence or another factor (such as an error or the presence of an external heating source). This is expected, since a low set temperature indicates the resident consciously "turns the heating off" and a low measure temperature indicates the house is cooled down after the central heating is turned off.

Table 7: Five days without gas_ch usage of heater_id 24137

heater_id	date_day	avg(t_set)	avg(t_act)	temp_diff	gas_ch	gas_dhw
24137	20191220	16.0	20.2	14.33	0.0	0.41
24137	20161225	14.0	17.7	12.11	0.0	0.13
24137	20190228	18.9	20.7	16.69	0.0	0.35
24137	20190227	19.0	22.1	16.76	0.0	0.37
24137	20190228	17.5	21.8	16.47	0.0	0.61

Classification

In order to work towards a classification model that determines whether or not a resident was absent during a certain day, heater id's with the most observations for the months January 2020 and December 2019 were used. After inspection, it was decided the month December 2019 had

the most interesting heaters to classify manually, since these heaters contained more variation in t_set and t_act than those in January 2020. Variation was necessary, since having very similar observations meant there was a lower chance of identifying absence at all. For example, if all observations contained a t_set of 19.0, it is most likely these observations did not contain absent days and are therefore not useful for determining what the maximum t_set when someone is absent should be.

Next, from the heater id's that contained different values for t_set, it was decided what value would be the maximum set temperature when the resident was absent. The average gas usage of the observations below this maximum was calculated and compared to the average gas usage of the observations above this maximum set temperature. It is expected that the average gas usage is lower when the t_set is lower. Some of the heaters that met this condition were taken for the first attempt at classification (e.g. 186773, 67171, and 77505). After an attempt at classifying 186773 (see table 8), a new question was raised: "Is it possible to distinguish absence for 24 hours from partial absence, such as coming home after work?". For example, the resident from table 8 was absent on the 7th of December 2019 according to our classification, because the set temperature was below 12. The actual measured temperature was 14.8 degrees Celsius this day. On the 22nd of December, the resident was present, but the measured temperature inside the home was only 15.3 degrees Celsius, which does not differ much from the 7th. However, since the set temperature is 19.0, the day was simply classified as present.

Table 8: First attempt at classifying heater 186773

heater_id	date_day	avg(t_act)	avg(t_set)	gas_ch	gas_dhw	avg(temp)/10	absent
186773	20191225	15,47	19,00	1,85	0,2899	6,71	No
186773	20191230	17,28	19,00	4,81	1,3536	4,52	No
186773	20191215	12	11,00	3,67	1,4514	7,11	Yes
186773	20191222	15,27	19,00	2,15	1,164	7,54	No
186773	20191204	13,65	9,00	0	0,8264	1,22	Yes
186773	20191216	12,41	9,50	7,74	0,8868	7,11	Yes
186773	20191217	14,07	11,50	3,19	1,2429	10,75	Yes
186773	20191201	11,64	9,00	0	0,8253	-0,76	Yes
186773	20191221	15,74	19,00	1,42	0,4717	7,59	No
186773	20191207	14,8	9,00	0	0,3193	8,78	Yes

The hypothesis is that some of these days can be classified as partially absent, since the average measured temperature differs quite a bit from the set temperature, which could potentially be explained by a delay in setting the higher temperature. Imagine this resident has the thermostat on approximately 15 degrees for the hours he is asleep and at work and turns the thermostat up to 21 degrees when he is at home after work. The average t_set would then be 19 degrees, indicating the resident was present throughout the 24 hours.

To counter this, a new column was added to the data frame containing the difference between the average actual temperature and the average set temperature, called thermodiff. This information was then used to create a new condition. Next to this, the new classification needed to filter out days with a relatively normal or high gas_ch values, since we expect a lower than average gas usage on absent days. Eventually, the classification was defined as follows:

If t_set < x & gas_ch < y, the resident was absent.
If t_set > x & thermodiff > z, the resident was partially absent.
If t_set > x & thermodiff < z, the resident was present.

After successfully classifying heaters 186773, 67171, and 77505 again with this definition, the next step towards testing the classification was made. In order to use the above classification on the entire data set in future research, it is necessary to create formulas for the values x, y and z. However, due to the limited number of observations in the current data frame it is challenging to create and verify the accuracy of the formulas. Next to this, the exploratory analysis proved it would be difficult to create such general rules as every resident's behavior is unique. Therefore, new rules were created and tested. In order to measure the performance of this classification, the amount of days classified as *absent* was compared to the expected amount according to CBS statistics (CBS, 2021). The new conditions, result and test of the classification described above are discussed in the next section.

Detecting Resident's Absence: Results

During this research, three different classifications were done. The first, and simplest, classification is discussed above. Next, two new methods were tested which will be discussed in this chapter. These include the manual classification using a condition for gas_ch and a rule that is able to check for partial absence and a method that is adapted for multiple heater ID's.

Manual classification

Heater 186773

Because this analysis was done on a sample, heater ID 186773 had a mere ten observations for December 2019. Although this is a low amount, it was still one of the heaters with the most observations for this month. Furthermore, there was a significant difference between set temperatures, making this heater interesting to classify (see table 8). First, the observations were manually inspected to select possible values for x (the limit for t_set). As mentioned before, the average gas usage of the observations that fall below x and above x were calculated and compared. In this case, 12 degree Celsius seemed a logical option. The t_set values were either below 12, which is considered an "off-temperature" (meaning it is a common temperature to insert when leaving the house), or 19 degree Celsius, which is a value where the heater is considered to be on. If t_set had a value above 12, it was assumed the resident was at home. Next to this, the gas_ch values had to be below average, since we expect a lower gas usage when someone is absent. After inspecting table 8, there was an attempt to classify some days as partially absent, due to the significant difference in temperatures. Eventually, the following rules were used:

If t_set <12 & gas_ch < mean(gas_ch) = Absent
If t_set >12 OR gas_ch > mean(gas_ch) & Δt_act/t_set >3 = Partially absent
If t_set >12 & Δt_act/t_set <=3 = Present

This resulted in the classification as can be seen in table 9. The days that have been classified as *absent* seem accurate, due to the low values for the set and actual temperature and 0 m^3 gas usage (ch). However, some days that are classified as *present* have low actual and set temperatures and it seems unlikely the resident was present (all day) at these uncomfortable temperatures. This is due to the added condition of gas_ch < mean(gas_ch) for absence. Since the coding ends with "otherwise, 'Present'", it now classifies low set temperatures with above average gas_ch usage as *present.* This is not desirable, and therefore, another revision of the conditions is needed. A final attempt at defining accurate conditions (or in this case, filters) will be addressed in the classification in a further paragraph.

Table 9: Manual classification of heater ID 186773 for December 2019

heater_id	date_day	avg(t_act)	avg(t_set)	gas_ch	gas_dhw	avg(temp)/10	absent
186773	20191225	15,47	19	1,85	0,29	6,7	Partially absent
186773	20191230	17,28	19	4,81	1,35	4,5	Present
186773	20191215	12	11	3,67	1,45	7,1	Present
186773	20191222	15,27	19	2,15	1,16	7,5	Partially absent
186773	20191204	13,65	9	0	0,83	1,2	Absent
186773	20191216	12,41	9,5	7,74	0,89	7,1	Present
186773	20191217	14,07	11,5	3,19	1,24	10,8	Present
186773	20191201	11,64	9	0	0,83	-0,8	Absent
186773	20191221	15,74	19	1,42	0,47	7,6	Partially absent
186773	20191207	14,80	9	0	0,32	8,8	Absent

Heater 67171

Next, heater ID 67171 was classified using the same rules, but with different values for x and y (x = 14 and y = 3.86). The value for z remained 3, since this temperature difference was still appropriate considering the distribution of temperatures. The final classification shows this heater ID did not have any days that meet the condition for *partially absent* (see table 10). In this case, the conditions seem to give accurate results, as the days that are classified as *absence* have lower than average values for t_act, t_set, and gas_ch. In order to test the applicability of the rule, a few different months and years were classified as well (see table 11 and 12). During this analysis, two interesting points were found: (1) the temperature limit for December in other years was equal to December 2019 and (2) the t_set value for February was 3 degrees Celsius higher (17.0). This proved it is important to consider the limited applicability of the rule in other seasons, where temperatures are higher and differences between outside and inside, and set and actual temperatures might be smaller.

Table 10: Manual classification of heater ID 67171 for December 2019

heater_id	date_day	avg(t_act)	avg(t_set)	gas_ch	gas_dhw	avg(temp)/10	absent
67171	20191228	15,2	14	0	0	-0,1	Absent
67171	20191209	20,1	17,7	11,85	0,94	6,2	Present
67171	20191203	20,7	18,7	8,74	0,92	3,8	Present
67171	20191219	18,4	16,5	1,48	0,02	10,6	Present
67171	20191213	15,2	13,5	0	0	4,3	Absent
67171	20191205	15,9	13	0	0	1,5	Absent
67171	20191201	17,8	17,5	9,62	0,02	0,4	Present
67171	20191229	14,2	12,7	0,12	0	0,5	Absent
67171	20191210	19,1	15,5	3,00	0,04	4,0	Present

Table 11: Manual classification of heater ID 67171 for December 2017

heater_id	date_day	avg(t_act)	avg(t_set)	gas_ch	gas_dhw	avg(temp)/10	absent
67171	20171226	16,9	14	5,04	6,0	5,6	Absent
67171	20171218	20,8	21,3	13,08	0,85	5,2	Present
67171	20171201	19,8	20,0	5,91	0,03	0,4	Present
67171	20171221	19,8	20,0	5,76	0,09	8,4	Present

Table 12: Manual classification of heater ID 67171 for February 2018

heater_id	date_day	avg(t_act)	avg(t_set)	gas_ch	gas_dhw	avg(temp)/10	absent
67171	20180224	16,2	15	0,20	0	-1,1	Absent
67171	20180210	17,0	16,7	0,15	0,05	2,5	Absent
67171	20180204	17,2	19,0	12,36	0	0,7	Present
67171	20180205	20,7	22	12,98	0,76	0	Present
67171	20180208	19,8	21	9,81	0,48	-2,3	Present

General rules for future classification

Although the classification of heater ID 67171 seems more accurate than 186773, it is not possible to apply these conditions on different heaters and periods. Until now the analysis and results were limited to within heater classification. Thus, next to improving the rule on gas usage it is also necessary to explore the possibilities of creating rules that are applicable on a larger sample of heaters. In order to assess what rules would be fitting, four rules were tested and compared. First, limits for t_set and gas_ch were decided by taking the average and decrease the value with *x* times the standard deviation. Next, the days that were filtered by this rule were counted and divided by the total amount of days (or total amount of rows = 707,468). This produced a percentage that was comparable to the CBS percentage.

T_set minus 2 standard deviation & gas_ch minus 1 standard deviation
If t_set < 11,986 & gas_ch < 2,534
1,2 %

T_set minus 1.5 standard deviation & gas_ch minus 1 standard deviation
If t_set < 13,3394 & gas_ch < 2,534
1,9 %

T_set minus 1 standard deviation & gas_ch minus 1 standard deviation
If t_set < 14,691 & gas_ch < 2,534
2,9 %

T_set is below average & difference between t_act and t_set is between -1,5 and 1,5 & total_gas_m2 below average
x = mean(t_set) = 17,396
y = mean(total_gas_m2) = 0,0684
If t_set < x & thermodiff > -1,5 & thermodiff < 1,5 & total_gas_m2 < y
6,5 %

These lines of code filter days from the original data frame that would be considered absent days. To have a quick overview of the amount of days, the percentage is calculated. For example, using the first rule approximately 1,2 percent of the days were absent days, meaning the resident was absent for 24 hours. After trying the first three rules, it was clear the difference in t_set was not impactful enough. The focus then laid on the gas usage. The filtering was done on a dataframe of December, January and February of a sample of different heaters, instead of within heater as before. So, due to a larger variety in gas consumption, setting a limit for gas

usage might have been too general of a rule. Therefore, the last rule made use of the column 'total_gas_m2'.

To assess which one of these percentages is accurate, statistics on Dutch holidays were consulted. According to the Dutch Central Office of Statistics (CBS), 83 percent of the Dutch people went on a holiday in 2019. The average Dutch person is on a holiday for 20,8 days of the year (CBS, 2021). So, we expect 5,7 percent of the days in a year to be classified as absence due to going on vacation. The last rule that considers total_gas_m2 generates a result which is comparable to the CBS statistics numbers.

Discussion

Base model

This research aimed to find the answer to the question "Can we build a base model to predict gas usage based on Intergas' data?". This indeed seems possible. In the process of this research, a random forest regressor model was used to predict resident gas usage. Although this model did not predict quite accurately, the results give a hopeful depiction of what is possible with data and gas usage prediction. Because this model had no baseline or reference point, it is hard to classify its performance as good or bad. Despite this baseline absence, the explained variance gives an indication of its usability and the predictors can be compared among each other.

The regressor model had an explained variance of almost 50 percent. This means 50 percent of the variance in the outcome could be explained by the features in the model, which is interpreted as high. Although these predictors seem to explain quite some variance, there is also more than 50 percent non explained variance. This is according to what was expected, as there are a lot of factors that influence one's gas usage. Next to this, households and their energy consumption behavior differ a lot from each other, which makes it highly complex to predict accurately.

For the created regressor model for this research, the feature importance can give insight to which features have a substantial impact on the prediction accuracy. Two features scored high on average, which were year and sun. According to the correlations (see figure 3), sun was expected to be a substantial predictor. Sun is, of course, highly correlated with temperature. This temperature feature did not have a substantial influence on the prediction. Random Forest is a model that is known for its way of dealing well with collinearity, which was the main consideration when selecting this model. This way of dealing with collinearity is a possible explanation for the high influence of sun and lower influence of temperature. Because they are collinear, they cannot both have a high influence. A more remarkable observation was the influence of year on the prediction within the model. Year is a derived variable from the observation timestamp. Observation year could have an influence because some years will have a lower temperature on average than others. This will probably influence the mean gas usage.

One of the main limitations of this research is the varying amount of records for the year values. Every ID's year value is the average of all the 24 hour values from this ID, within a certain year. This varying amount of records could have an influence on the results. For example, for some heater ID's only data points from the winter period are missing. This would give biased estimates and has influence on the averages that were used to train the model. A proposed solution could be to make a distribution to see which values are missing, if for instance the missing values are in a heating curve the slope value of this curve could be used to calculate the missing value.

The random forest regressor model has the property of being an easily applicable model, which often works well, also without any parameter tuning. This property helped to overcome the problem of complex steps to make an algorithm work. At the same time, the prediction accuracy could probably be improved by tuning these parameters. This parameter tuning could be a good followup research topic.

Detecting resident's absence

The aim of the second part of the thesis was to explore the possibility of correcting for resident's behavior. In this particular case, the research focused on detecting resident's 24 hour absence with the available data. From this limited research, it is concluded that it is possible to classify for absence using the heater data. However, selecting values for analysis and creating rules and conditions was often done by trial and error. For future research, a more systematic approach is preferred. For example, a method similar to the method of filtering and comparing percentages to available statistics could be developed and tested on a larger data set.

Next to this, it may be possible to classify partially absent days. To confirm this, more research would be needed into the assumptions made. For example, studies that confirm certain patterns of behavior on partially absent days (such as working days) could be used to validate the results from data analysis.

Lastly, this research aimed to explore what variables are important to take into consideration when trying to detect (partial) absence. It was concluded the following variables are leading in creating conditions for classification: the set temperature (t_set), the actual temperature inside (t_act), the gas usage of the central heating (gas_ch), the gas usage per square meter (total_gas_m2) and the difference between t_act and t_set (thermodiff).

A limitation of using these variables, especially temperature variables, is that the current conditions used are most likely only applicable for the winter months (December, January and February).[3] Since some assumptions rely on a certain distribution of temperatures and temperature differences, and these are not necessarily similar in summer, this particular classification is only reliable in the winter months. For example, in the summer the gas usage is much more often lower than average and close to zero. Gas_ch is not a reliable variable during the summer for classifying absence, as a low gas usage is expected with higher outside temperatures, whether the resident is absent or present. If colder autumn and spring days need to be classified, different conditions for classification would be needed. For example, the classification could still be done by comparing values for gas usage.

Further limitations of this research include the scope of the project, the data frame used and the insufficient validation of results. Due to time constraints it was not possible to analyse and validate more heaters under different circumstances. For example, it is advisable to analyse

[3] Note that February was only tested for one heater ID, so it is unclear whether or not February has the same accuracy in results compared to the colder December and January months.

more samples, or even better, the entire data set. Next to this, it would be possible to use an unfiltered data frame to explore differences in results.

Another interesting point for future research would be to look into more studies regarding the average absence of Dutch people. For example, the CBS statistic of 20,8 vacation days that was used to validate the accuracy of the filter rules may not have been the most accurate choice. Since vacation days are not distributed equally throughout the year, but are rather concentrated around different periods (e.g. school holidays), it is not possible to assume each resident has an expected absence of approximately 5,7 percent for each month.

All in all, it is possible to classify for absence using the heater data from Intergas. How accurate these classifications are depend on the conditions used and how these were established. It also seems possible to create rules that are applicable for multiple heater ID's. If these approaches are improved and systematically applied, it may be possible to correct for resident's absence in the base model. However, resident's absence is one of many factors influencing gas consumption and is only one part of resident's behavior. Further research should focus on how to detect other types of behavior. The next step would be to find ways for correcting absence and these other behaviors to eventually improve the prediction model.

Bibliography

Berkland, S. M. (2014). A Comparison of American, Canadian, and European Home Energy Performance in Heating Dominated–Moist Climates Based on Building Codes.

Boonekamp, P. G. (2007). Price elasticities, policy measures and actual developments in household energy consumption–A bottom up analysis for the Netherlands. *Energy Economics*, *29*(2), 133-157.

Brom, P. van den, Meijer, A., & Visscher, H. (2018). Performance gaps in energy consumption: household groups and building characteristics. *Building Research & Information*, *46*(1), 54-70.

Centraal Bureau Statistiek (CBS). (2021). *Vakanties van Nederlanders; kerncijfers*. <https://www.cbs.nl/nl-nl/cijfers/detail/84363NED>.

Energielabel. (n.d.). *Veelgestelde vragen*. <https://www.energielabel.nl/woningen/veelgestelde-vragen/>.

European Commission. (2020, 17 February). *In focus: Energy efficiency in buildings.* <https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-feb-17_en/>.

European commission. (n.d.). *2050 long term strategy.* <www.ec.europa.eu/clima/policies/strategies/2050_en/#tab-0-0>.

Feenstra. (2019). *Sluipverbruik van stroom en gas.* <https://www.feenstra.com/zorgelooswonen/sluipverbruik-van-stroom-en-gas/>.

Honold, A. Towards Data Science. (2020). *The What, Why, and When of Apache Spark.* <https://towardsdatascience.com/the-what-why-and-when-of-apache-spark-6c27abc19527>.

Intergas. (n.d.). *Over Intergas.* <https://www.intergas-verwarming.nl/consument/over-intergas/>.

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

Martens, S., & Spaargaren, G. (2005). The politics of sustainable consumption: the case of the Netherlands. *Sustainability: science, practice and policy*, *1*(1), 29-42.

Majcen, D., Itard, L. C. M., & Visscher, H. (2012). Theoretical vs. actual energy consumption of labelled dwellings in the Netherlands: Discrepancies and policy implications. *Energy policy*, *54*, 125-136.

Majcen, D., & Itard, L. (2014a). Relatie tussen energielabel, werkelijk energiegebruik en CO2-uitstoot van Amsterdamse corporatiewoningen. *Delft University of Technology (OTB): Rekenkamer Metropool Amsterdam.* <http://resolver. tudelft. nl/uuid: b0b73c48-4413-4dda-8b1b-748cf65a534b>.

Majcen, D., & Itard, L. C. M. (2014b). Relatie tussen huishoudenskenmerken en-gedrag, energielabel en werkelijk energiegebruik in Amsterdamse corporatiewoningen.

Majcen, D., Itard, L., & Visscher, H. (2015). Statistical model of the heating prediction gap in Dutch dwellings: Relative importance of building, household and behavioural characteristics. *Energy and Buildings*, *105*, 43-59.

Radar. (2021, 3 March). *Tot 500 euro voor het energielabel, maar wat heb je eraan?* <https://radar.avrotros.nl/uitzendingen/gemist/item/tot-500-euro-voor-het-energielabel-maar-wat-heb-je-eraan/>.

RvO. (2020a, March). *Opnameformulier behorend bij het opnameprotocol.* <https://www.RvO.nl/sites/default/files/2020/10/opnameformulier-behorend-bij-het-opnameprotocol-nta-8800-versie-maart-2020.pdf>.

RvO. (2020b, 29 May). *Energielabels op basis van NTA 8800 bij bouwaanvraag EPC.* <https://www.RvO.nl/sites/default/files/2020/06/energielabels-op-basis-van-nta-8800-bij-bouwaanvraag-epc.pdf>.

RvO. (2021, 29 June). *Energielabel woningen.* <https://www.RvO.nl/onderwerpen/duurzaam-ondernemen/gebouwen/wetten-en-regels/bestaande-bouw/energielabel-woningen>.

Santin, O. G. (2011). Behavioural patterns and user profiles related to energy consumption for heating. *Energy and Buildings*, *43*(10), 2662-2672.

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

Yunus A. Çengel, & Ghajar, A. J. (2020). *Heat and Mass Transfer: Fundamentals [and] Applications*. McGraw-Hill Education.

Appendix A: Base Model Code

Appendix B: Classification Code