Doing More with Less - 1

Master Thesis

Maria Fakou

Utrecht, July 1, 2022

Information Page Graduation Report

Utrecht University
Heidelberglaan 8, 3584 CS Utrecht, Netherlands

Master Thesis

Name of student:	Maria Fakou
Student number:	4829271
Course:	Master Applied Data Science, Utrecht University
Period:	April 2022 - June 2022
Company name:	Intergas Verwarming B.V.
Address:	Europark Allee 2
Postcode, City:	7742 NA Coevorden
Country:	Netherlands
Company supervisors:	Erwin Bisschop
Email:	erwin@inversable.com
University supervisors:	Arno Siebes
Email:	a.p.j.m.siebes@uu.nl
First Examinator:	Arno Siebes
Second Examinator:	Ad Feelders
Non-disclosure agreement:	Yes
Number of words:	8672

NDA and Working Arrangements

for Students working with the Intergas Data

1. You have your own clean spark environment (no other users are using it). After your research is done, we will destroy your access to the environment (but results will be kept on the server).
2. Do not copy data from the server to your local device but do all processing on server. Datasets are intellectual property of Intergas.
3. Outcomes (not raw data) may be downloaded to your own pc for the report.
4. There are no access limitations to folders on the server / filesystem, but try not to delete files which you will need yourself. If something is gone, it is gone.
5. Make sure you have discussed with Intergas about which findings you may / may not publish. By default you cannot publish anything you discover in the data without consent from Intergas. So make sure Intergas has enough time to give you feedback on your (semi-)final version of your thesis.
6. Do not use the server for things that are not a part of your data assignment.
7. Do not provide access to the server to others outside of your group.
8. Never distribute, copy, sell or share the data of Intergas and make the necessary precautions to prevent this from (accidentally) happening. For example, store your access token in a secure way.

I, Maria Fakou (full name),

have read the agreement and will stick to it.

Date: 04/28/2022

Signature: Maria Fakou

Abstract

The aim of this study is to find out from what point in time and with what amount and type of data you can detect with a certain amount of certainty a significant decrease of the gas consumption for an individual household. Data points for the summed gas consumption for the average temperature differences between indoor and outdoor temperature for each day for annual periods between September and April from 2015 till 2020 were taken. To be able to make the earliest possible detection of a valid decrease of gas consumption, three consecutive heating periods are needed.

Afterwards, the slopes were compared with the following period slopes to identify an increase or decrease. If there is a significant change that was determined differently in three different approaches, you can assume that a possible reason is a newly add insulation of that household. Those household where a significant decrease has been detected by the different approaches linear regression, Support Vector Regression and Random Forest, were afterwards filtered out to have a final dataset with houses where an insulation has possibly been added.

The findings of the study showed that with two linear models, linear regression and support vector regression, significant decreases in gas consumption can be detected in the data.

These results lead to the assumption that the gas consumption and the average temperature difference per day alone show a change in gas consumption, but this cannot be attributed to a newly added insulation, as this can also have many other reasons.

Preface

The study was done as part of a task for the company Intergas Verwaming BV and is divided into three main tasks. The first part is a collaboration between the three applied data science students from Utrecht University, in which they prepare the data provided by Intergas. The aim of this work is to create a data set that is as meaningful as possible and as close to reality as possible in order to learn and test different models for use.

In the second task of this thesis, each of the three students works individually on the model for the gas use. Varoon Sushil Agrawal is working on processing the various slopes for each heater in the prepared data with linear regression to detect significant changes. Maria Fakou researches with a random forest regression model to find a different way of detecting changes and Moritz Münten applies a Support Vector Regression model for calculating the slopes and detecting significant decreases.

The third and final part of this study is again a joint comparison of the different results in order to make assumptions about which model is most suitable in the context of the task. Here the three students come to a common conclusion about the study, answer the research question and make a recommendation for further research.

Statement of Authenticity

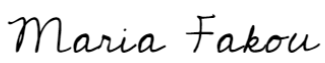
I, the undersigned, hereby certify that I have compiled and written the attached document / piece of work and the underlying work without assistance from anyone except the specifically assigned academic supervisors and examiners. This work is solely my own, and I am solely responsible for the content, organization, and making of this document / piece of work.

I hereby acknowledge that I have read the instructions for preparation and submission of documents / pieces of work provided by my course / my academic institution, and I understand that this document / piece of work will not be accepted for evaluation or for the award of academic credits if it is determined that it has not been prepared in compliance with those instructions and this statement of authenticity.

I further certify that I did not commit plagiarism, did neither take over nor paraphrase (digital or printed, translated or original) material (e.g. ideas, data, pieces of text, figures, diagrams, tables, recordings, videos, code, ...) produced by others without correct and complete citation and correct and complete reference of the source(s). I understand that this document / piece of work and the underlying work will not be accepted for evaluation or for the award of academic credits if it is determined that it embodies plagiarism.

Name:	Maria Fakou
Student Number:	4829271
Place/Date:	Dilkrath, July 1, 2022

Signature:



Contents

List of Figures

List of Tables

1 Introduction

In the fight against the climate crisis, one tool is to drastically minimize our energy and gas consumption. The housing sector is a huge consumer of the energy and plays a vital role in achieving energy efficiency targets in the EU (Faidra Filippidou 2018). Due to poor energy performance of buildings, they account for 38% of total energy consumption in the European Union(EU) (Delft CE 2015). Out of which, households are responsible for 24.8% of final energy consumption in the EU (*Consumption of energy* 2016). Thermal comfort in housing is established by space heating by maintaining the indoor temperature at a desired, uniform level and providing proper admission of fresh air (Haris Lulic 2013). In the Netherlands, 85% of the households are heated using natural gas (Faidra Filippidou 2018). So to contribute to solving the challenges of the climate crisis, one first step is to reduce the energy and therfore gas consumption of individual households in the Netherlands.

Intergas Verwarming BV. builds and sells heating equipment from gas boilers, water heaters, hybrids and control devices to heat pumps. Through various contracts with their customers, Intergas has a detailed, large accumulation of data of the respective energy use of their clients. However, at the current time of rising energy prices and inflation, energy consumption by individual households is also becoming increasingly expensive. Of these, many consumers and landlords are already deciding to build their properties energy poorer and to insulate them better afterwards. Intergas is already exploring different ways to identify these newly built houses based on their data in order to better manage their energy budget through houses that have been newly installed and therefore consume less energy. Intergas also want to share this information with their customers to show them the benefits of a new insulation, which is a possible percentage decrease in gas consumption so that they can save costs.

There are now two essential challenges. On the one hand, Intergas would like to know how quickly and with what certainty one can say something about the changes in energy consumption. This is about the temporal aspect as well as the data aspect, because you collect data over a certain period of time, but you want to know with what amount of data you can say something about the changes with certainty. Secondly, how certain is the change in slope associated with a change in insulation? In this context, slopes are the increasing summed gas consumption from an individual heater per temperature difference of inside and outside temperature. After calculating the differences after a new insulation, it becomes clear that these only become apparent at a higher energy consumption, which is usually the case when temperatures are colder than in summer when heating is hardly used.

Thus, the main question of this research: How soon can we say something about a new slope with certain amount of certainty?

First, the data made available must be processed and then used for the models. With a data exploration analysis is trying to find out how many data points are needed to calculate a statistically relevant slope can be drawn for the consumption of the gas. Additionally, whether these data points are compared on a daily monthly or periodic basis. After differences are calculated with the various changes, an attempt is made to detect significant changes by adjusted filter functions and by comparing increased error rates in a prediction model.

Finally, the aim of this research is to compare the different results of the detection of a significant decrease in gas consumption. And classify this difference whether it is due to a newly added insulation of the individual household.

2 | Data

The data were provided by Intergas to perform the current analysis. To gather all the needed information the following four datasets were combined.

Column Name	Type	Description
heater_id	Integer	Heater unique identification number
gas_use	Double	Gas consumption in m^3/hour
surface_area	Integer	Surface area of the house in m^2
t_set	Double	Temperature set on the thermometer (C)
t_act	Double	House temperature (C)
TimeKey	Timestamp	year/month/day hour

Table 2.1: Ig_gasuse_hourly.

Column Name	Type	Description
HEATER_ID	Integer	Heater unique identification number
wijk	Integer	Neighborhood
building_year	Integer	Building year

Table 2.2: ig-heater-info-nl-2.

Column Name	Type	Description
wijk	Integer	Neighborhood
rain	Double	Rainfall amount in 0.1 mm
sun	Double	Amount of sun in 0.1 hours
temp	Double	Temperature (C) * 10
wind	Double	Wind in 0.1 meters/second
TimeKey	Timestamp	year/month/day hour

Table 2.3: od_knmi_hourly_wijken_v2.

Column Name	Type	Description
HEATER_ID	Integer	Heater unique identification number
WONING_TYPE	String	House type

Table 2.4: House_prop.

2.1 Data preprocessing

In the first stage of the data preprocessing, it was considered of paramount importance to inspect the datasets individually and delete problematic values to reduce their size and the computational time of the analysis, but also to improve the quality of the results. Following are the steps taken:

- The data recorded from May until August were removed, since the gas consumption during these months is negligible for heating. This operation was applied to *Ig_gasuse_hourly* and *od_knmi_hourly_wijken_v2*.
- Buildings of size below 40 or above 400 square meters, in *Ig_gasuse_hourly*, were filtered out, as they do not provide any useful information to the current research.
- The upper threshold of 26 and lower threshold of 0 degrees Celsius was set for t_set, while the upper threshold of 30 and lower threshold of 10 degrees Celsius was set for the t_act, in *Ig_gasuse_hourly*. The remainder of the records is assumed unlikely to be accurate.
- Heaters that did not have building year or neighborhood were removed from *ig-heater-info-nl-2*.
- Houses that had a missing house type in *house_prop* were discarded.
- The minimum building year was 1005 and 25% of the values fell before 1956, hence it was decided to delete these data from *ig-heater-info-nl-2*, as they were odd. Specifically, the research was limited to buildings constructed from 1950 onwards.

To result in the final dataset left inner joins were performed to select the records that match in both datasets and prevent missingness of information. The datasets were joined as shown in table 2.5.

Left table	**Right table**	**Key**	**Table Name**
Ig_gasuse_hourly	ig-heater-info-nl-2	heater_id	Join_1
Join_1	od_knmi_hourly_wijken_v2	Wijk, TimeKey	Join_2
Join_2	House_prop	heater_id	Final_df

Table 2.5: Left inner join tables.

Consequently, duplicate rows were detected and deleted, as well as records of the same house and timestamp that contained different measurements for the gas usage or the inside temperature. In the latter case, every record related to these heaters was removed and considered incorrect. Heaters monitored for a single period were also removed from the dataset. A period includes data for the months September to April, under the hypothesis that insulation is mostly added during the summer months. Hence, if there is a shift to be detected, it will be between these heating periods, and not between calendar years.

Additionally, the following table describes the datasets size before and after the related filters.

Dataset	Before filtering	After filtering	Percentage removed
Ig_gasuse_hourly	558,960,694	354,261,532	36.6%
ig-heater-info-nl-2	39,305	39,175	0.33%
od_knmi_hourly_wijken_v2	74,894,318	51,603,006	31.09%
House_prop	39,305	39,155	0.38%
Final_df	324,849,444	222,216,880	31.6%

Table 2.6: Datasets size before and after filtering

For further preparation of the data, the outside temperature was divided by 10 and was subtracted from the indoor temperature (t_act - temp). The resulting difference denoted the insulation level of the house and was a determinant variable of the research objective, namely, to identify the change in energy consumption by early detection of improvement in house insulation. Negative values of this difference were not reliable; thus, these data were removed.

Insulation directly affects gas use, so the temperature difference could be used to build a simple and quite accurate model, without including the variables of weather conditions. Moreover, zero gas use during some hours of the day implied better predictions for daily data than for hourly data. As the hourly values could adversely affect the regression models, the data were grouped by period, month and day of the month, summed by gas use and averaged by temperature difference.

The time information was extracted by the *TimeKey* timestamp and the heating periods were defined as presented in table 2.7:

ID	**Period**
1	Sept. 2015 - Apr. 2016
2	Sept. 2016 - Apr. 2017
3	Sept. 2017 - Apr. 2018
4	Sept. 2018 - Apr. 2019
5	Sept. 2019 - Apr. 2020

Table 2.7: Heating periods.

The structure of the final dataset and its first five rows are depicted in tables 2.8 and 2.9, respectively.

Column Name	**Type**
heater_id	Integer
period	Integer
month	Integer
dayOfMonth	Integer
sum_gas	Double
avg_t_diff	Double

Table 2.8: Final dataset structure.

ID	heater_id	period	month	dayOfMonth	sum_gas	avg_t_diff
0	93059	3	4	14	2.7573	10.622500
1	93059	4	10	9	1.6920	8.964167
2	96265	5	1	11	6.0406	15.012917
3	66595	2	3	11	6.4874	11.985000
4	54477	4	10	30	5.6728	15.618750

Table 2.9: First rows of the final dataset.

2.2 Exploratory Data Analysis

The dataset contains 6,886,234 records of 12,675 heaters from October 10th, 2015, until March 1st, 2020. The number of records of a heater was not necessarily equivalent to other heaters, meaning that some heaters were measured for longer periods than others. In addition, data from 308 heaters related to a single period were not valuable for this research.

Table 2.10 shows the descriptive statistics of the daily gas use and average temperature difference. Both the daily gas consumption and the temperature difference presented extreme values on some occasions, while their most common values, or medians, were 4.66 and 12.10, in the given order.

summary	sum_gas_use	avg_t_diff
mean	5.376	12.162
stddev	4.459	4.31
min	0.0	0.01
25%	1.723	8.912
50%	4.669	12.107
75%	7.821	14.99
max	74.567	33.44

Table 2.10: Descriptive Statistics.

To understand the relationship between these instrumental variables for the current exploration, the Pearson Correlation Coefficient was computed and its value of 0.6 revealed that the daily gas use and the temperature difference were positively correlated. As illustrated in figure 2.1, there was a moderately strong, positive, linear association with a few outliers. This association justified the choice of linear regression models, which considered as suitable to estimate the difference in gas use between every two sequential periods.

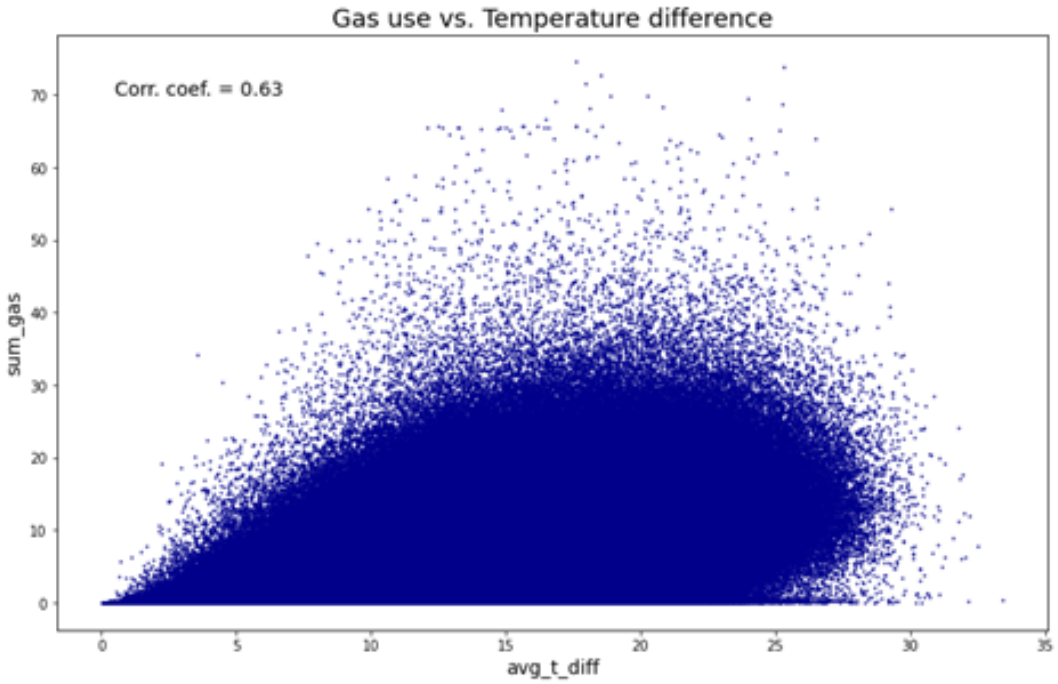


Figure 2.1: Gas use vs. temperature difference.

Furthermore, the 1st period contained the fewest data, namely the 5%, and the 2nd period consisted of the second smaller share of the dataset, the 14%. Data from the 4th period exceeded the rest, still those from the 5th and 3rd periods were nearly a quarter each, i.e., 25.2% and 24.1%, respectively. Therefore, the first period could not be perceived as a representative sample of the data, yet it was included in the three types of models, as the objective of this analysis was to test how fast a change can be detected using the least possible amount of data.

As expected, the gas consumption was higher during the winter months and decreased significantly in April, September, and October. The same trend was noticed for the temperature difference as well, while both cases suggested September to be the warmest month, as it had the lowest gas use and temperature differences *(Figure 2.2)*.On the other hand, no pattern was detected on the gas use or temperature difference during the separate days of the months, which was a reasonable inference, and indicated uniformity across the daily behavior of the users.

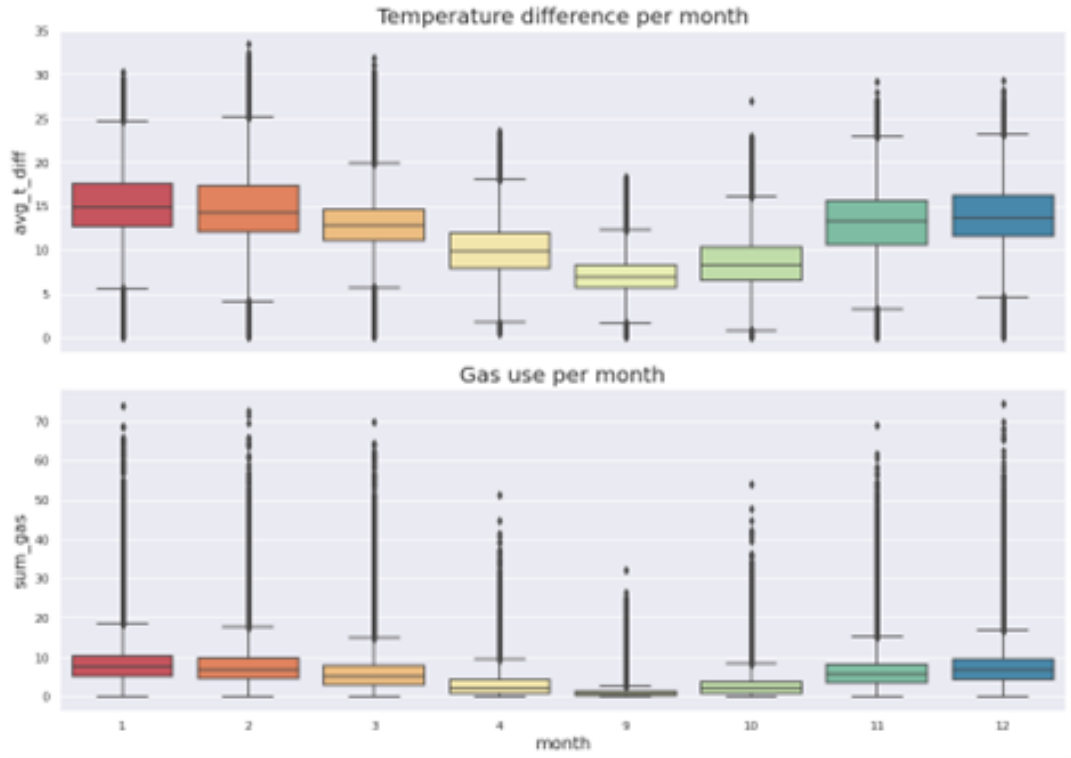


Figure 2.2: Temp. diff. and gas use per month.

Some examples of heaters were selected for further investigation, as the initial aim was to distinguish those that presented reduction in gas use, and then to examine how soon the distinction can be drawn. Figure 2.3, demonstrates three heaters of whom *8180* and *27729* were potential houses that added insulation during their recording by Intergas. Heater *8180* seemed to lower its gas use dramatically after the 1st period, whereas heater *27729* appeared to suddenly decrease after the 3rd period, and the gas use of both houses was stabilized immediately after declining. The gas use of *5924*, in contrast, remained quite stable trough the different periods and thus, it was assumed that the specific house did not improve its insulation level.

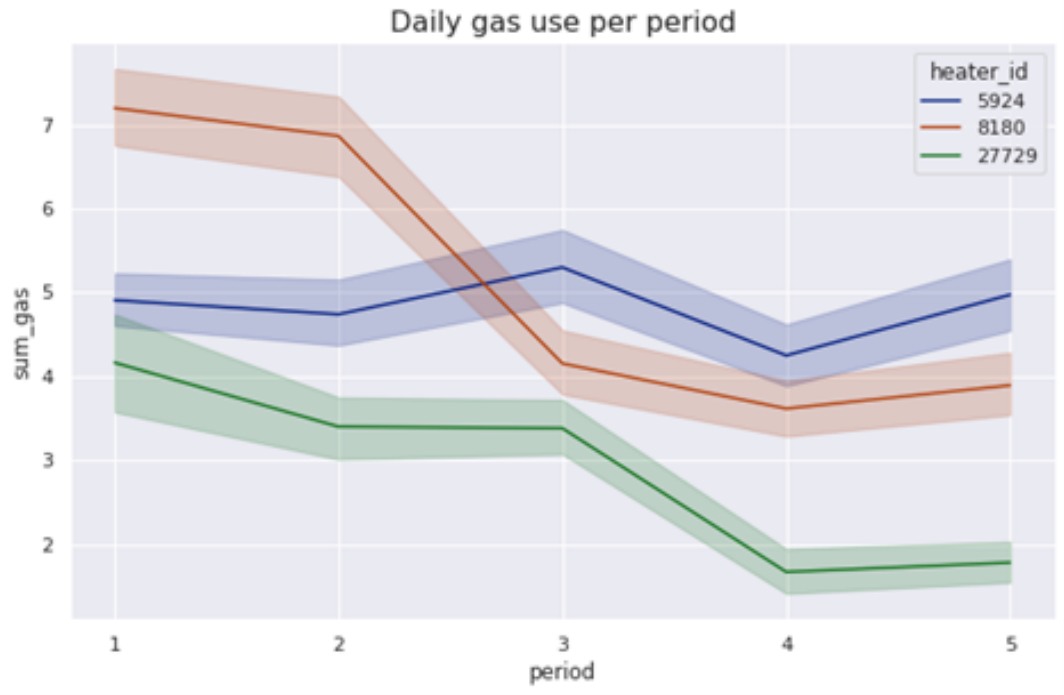


Figure 2.3: Daily gas use per period.

It is essential to highlight that these data were anonymous, meaning that they could not be connected to the individuals who own the heaters. Access to them was given by Intergas to recognize as soon as possible the decrease of their gas use caused by added insulation, but only data experts of the company would be able to interpret the location of the clients or their actual name.

3 | Methods

This research aims to help Intergas reduce the energy consumption of their clients by detecting the change in insulation and further, in gas use. To identify this change, three different regression models were trained: Simple Linear Regression (LR) (Agrawal 2022), Support Vector Regression (SVR) (Münten 2022), and Random Forest Regression (RF). The current thesis uses the RF model, which is by nature a nonlinear model

3.1 Random Forest Regression

Regression is a supervised technique, meaning that the model is trained on labeled data, i.e., known output, to estimate a relationship between the input and the output, or the independent and dependent variables respectively. The model's objective is to predict unseen data, or testing data, by learning some training data, and hence to be able to generalize and make accurate predictions on both sets. If the training data are not a representative sample of the population, the model captures this specific relationship and is unable to predict new data correctly. This problem is known as overfitting and can be avoided by applying appropriate settings to the model if the data do not allow further improvement.

Random Forest or Random Decision Forest is a collection of decision trees, suitable for regression and classification problems, and trained with the "bagging" (bootstrap aggregating) method. Bagging, which is an ensemble method, combines the predictions from multiple learning models to improve predictions and raise the stability of the final model, and since the predictions, in this case, were continuous outcomes, this problem could be solved using regression models. (Schonlau 2020)

For understanding how RF works it is helpful to look at the well-known Decision Tree. A decision tree can be considered as a set of the best questions someone might ask to a dataset so that a sample will be predicted as accurately as possible. Additionally, RF separates the data into multiple random subsets and generates a decision tree of each subset. Every decision tree predicts a result, and these various results are averaged and selected as the final prediction.

There are specific assets of RF that identify it as a remarkable model used for several regression and classification problems. For example, the model is non-parametric and resistant to outliers. Conversely, the algorithm is prone to overfitting and generally slower than simple linear regression, but the main complication for the present analysis was that the model cannot be interpreted by coefficients and intercept.

3.2 Hyperparameter Tuning

Even though RF is often used in complex datasets with multiple features, it could be applied in the current data on the condition that a small number of trees was chosen to form the forest. In that way, overfitting was prevented, and the present approach could be used in future research. Thus, hyperparameter tuning was applied to find out under which conditions the performance improves. Except for the number of trees, the maximum depth was taken into consideration, which indicates the longest path from the root to the deepest leaf node that a tree in the forest can have. Large values of depth can cause overfitting, and as only one feature was considered for the model's decisions, it was necessary to restrict the parameter space accordingly.

(Koehrsen 2018)

The hyperparameter space was adjusted according to the number of available training and testing data. Two types of experiments were conducted; models were trained for each period and each month of every period. The latter case included much fewer data points than the former one, and hence smaller parameter values were required to tune the models. Specifically, the models of every period were tested for larger parameter values, but they performed worse, and they were excluded from the analysis to boost their reliability and computational efficiency. The best values of the parameters were selected after an extensive search of specific parameter values. The candidate values of the parameters used to tune the RF models are presented in the tables below.

Parameter	**Parameter space**
Number of trees	[1, 2, 3]
Maximum depth	[1, 2, 3, 4, 5, 6]

Table 3.1: Candidate parameter values of models per period

Parameter	**Parameter space**
Number of trees	[1, 2, 3]
Maximum depth	[1, 2, 3, 4]

Table 3.2: Candidate parameter values of models per month

Each model was fine-tuned using the 3-fold cross-validation method, while the performance was measured and evaluated by the mean squared error. In more detail:

- The k-fold cross-validation technique is a resampling process that partitions the dataset into k sets. (Brownlee 2018) For each set, the specific set takes the role of the testing set and the rest of the sets compile the training set. In other words, each set is used one time as the testing set, k-1 times to train the model, and the summary of the k performances is returned to evaluate the model. For this analysis, three folds or sets were selected because of the small data size, as more partitions of the data would be unnecessary.
- The mean squared error (MSE) denotes the average squared difference between the true values and the predicted values, and therefore it is always positive. This metric uses square units, which render it easily influenced by large errors, and it was used as the evaluation criterion in the grid search to select the best combination of the hyperparameters. It is calculated by the following formula for a sample of n data points, where Y_i are the true values and $\hat{Y}_i$ the predicted values. (*Mean squared error* 2022)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

3.3 Procedure

After hyperparameter tuning, the models were trained, using the best values of their parameters, in two ways: for each period, and each month of every period. Each model was trained by the data of one period and evaluated by the data of the next heating period. The result was the predictions for both sets, which were

subtracted from the true values of the daily gas used to compute the prediction errors. Since the heating periods were five, models were trained for the first four periods, as the last period was solely used for evaluation, and only heaters that were recorded for more than one period were included.

While Random Forest is a complicated model, setting a threshold for the appropriate amount of data was cumbersome and since the research's objective was to do more with less, the single restriction was the minimum possible number of data points to perform any necessary estimation. Consequently, a model was trained only if there were available data from a certain heater for more than one period, and if there were at least 6 days recorded for each period. Less than that caused problems during hyperparameter tuning and metrics' computation, but it was also presumed that fewer data points would cause the models to learn the specific data points extremely well and fail to predict new unseen data.

Furthermore, the distributions of the prediction errors were investigated to reveal the degree of dissimilarity between the train and test sets. If the two distributions were homogeneous, it was beyond doubt that no change could be detected, as the model of the past month could fit the latest data and produce similar errors to the earliest data. On the contrary case, it was assumed that the train and test sets were not drawn from the same distribution, meaning that the daily gas use of the two consecutive periods was not commensurable, and there was a change in insulation.

To determine whether the two samples were similar t-tests and z-tests were applied to each set of training and testing prediction errors under the assumption that these residuals followed the Gaussian distribution. Z-tests (*Z-test* 2022) were performed in case the sets consisted of more than 50 data points and t-tests (*T-test* 2022) were used for some heaters measured for less than 50 days per period, to establish the significance of the similarity of the samples provided. The null hypothesis, that the two means were statistically similar, was rejected when the p-value of the test was smaller than the significance level of 1%.

To evaluate the predictions and compare the entire performance of the various models, the prediction errors were assessed by basic statistics as the mean and the standard deviation, and the quality of the estimators was quantified by the mean absolute error (MAE). (*Mean absolute error* 2022) The MAE is the averaged sum of the absolute prediction errors, as presented in the formula below, and it is an interpretable estimator of the model's predictions, because of its dependency on the measurement scale of the variable. For a given sample of n data points, where Y_i are the true values and $\hat{Y}_i$ the predicted values, the MAE is defined as:

$$MAE = \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|}{n}$$

Since the prediction errors were supposed to be normally distributed, to be able to distinguish whether there is a noticeable change, heaters that resulted in unusual values of MAE test score, namely outliers, were selected based on the three-sigma rule of thumb. The 3σ rule is also known as the 68–95–99.7 rule or empirical rule, and it is used to represent how the data lie within a normal distribution. (Hayes 2022) According to this rule, 99.7% of the data points lie within three standard deviations from the mean, and the rest 0.3% are considered outliers

4 | Results

First, the models were trained, for every period and heater individually, with data from 12,367 houses,to examine several independent clients and provide an accurate insight into the problem of detecting added insulation. The specific number of different heaters was regarded as suitable to appraise the following findings with a certain confidence, and the required computational time was feasible to conduct experiments, as it was estimated that every heater spent approximately 1,5 minutes for hyperparameter tuning and training.

During the analysis, various hyperparameters were tested for these models and it was noted that larger values of the max depth and number of trees yielded lower performance, especially in cases of low data quantity. These cases appeared to cause overfitting and generally larger prediction errors in both train and test sets. In addition, larger parameter values required higher computational time than the parameters mentioned in table 3.1.

To test whether the difference between the prediction errors of each pair of train and test sets were statistically significant, t-tests and z-tests were applied accordingly. The results exhibited that the residuals from 8478 heaters had different population means, while 3117 of them showed this inconsistency in more than one pair of heating seasons. An example of this odd case, including the mean of each set, is illustrated in figure 4.1, which explained why data from approximately 25% of heaters were determined as statistically different in more than two periods. As witnessed, the means of the two sets seemed to be more distant when one of the sets contained fewer data points, and it is important to mention that only the means of the top left and bottom right density plots were statistically different.

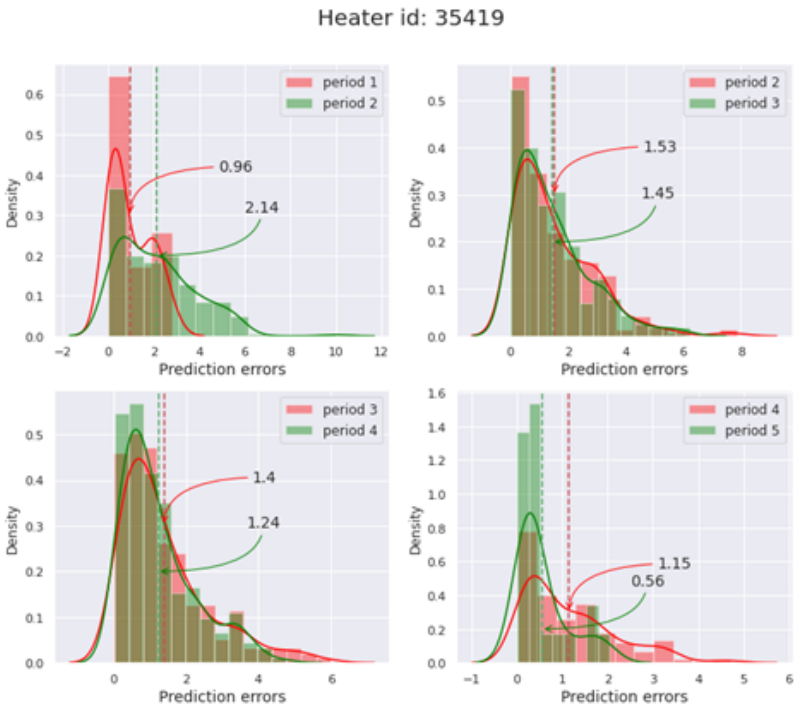


Figure 4.1: Heater 35419 - prediction errors per period.

Other cases reflected similar findings but also highlighted that these statistical tests could not extrapolate a conclusion for the actual change between the train and test set. As shown in figure 4.2, the pairs of error distributions were quite similar in the upper right and bottom left plots, although their means appeared to be statistically different. The p-value of the z-test conducted on the sets of the upper right plot was almost equal to the p-value of the bottom plot and higher than the p-value of the upper left populations, yet all of them were lower than 0.001 and the null hypothesis was rejected.

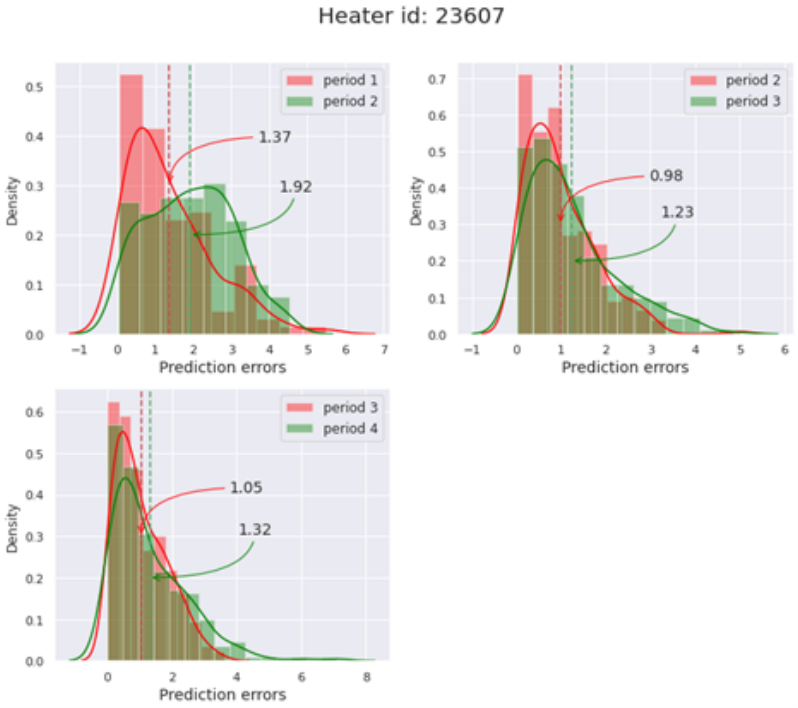


Figure 4.2: Heater 23607 - prediction errors per period.

Since it was impossible to compare each heater separately, the succeeding step was to observe the distribution of the evaluation metrics for each period, and even though the distributions per period were quite similar, the MAE was reviewed for each of them individually. This inspection was deemed essential to identify abnormal values, as most of the heaters did not have an equal amount of data per period and some heating seasons had generally fewer data, e.g., the first period. As presented in figure 4.3, the long right tails of the distributions suggested irregular values of the MAE in both sets and every period, while the range of the test scores was broader than the range of the train scores.

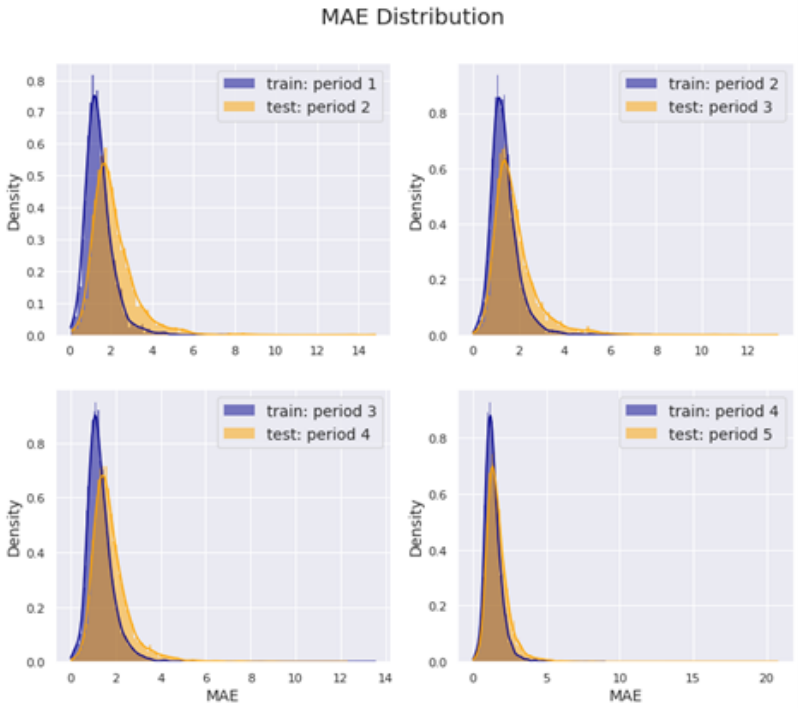


Figure 4.3: MAE Distribution per period.

Given that the prediction errors follow the Gaussian distribution, the three-sigma rule of thumb was applied to find out the heaters that generated these outliers. Specifically, the MAE of the train scores was filtered out to discover the values that fall into the range of 99.7%, and the MAE of the test scores was filtered out, in contrast, to find the outliers, i.e., the 0.3% of these data. In that way, models that performed well for the train set, but not the test set, could be differentiated from models that displayed deviant behavior, to wit, models that failed to make accurate predictions for data that were already known to them. Table 4.1 outlines the MAE mean and standard deviation of the data of every period, rounded to two decimal places, that were used to calculate the threshold of $\mu + 3\sigma$.

MAE		Mean (μ)	Standard deviation (σ)
Period 1	Train	1.45	0.67
	Test	2.13	1.07
Period 2	Train	1.38	0.62
	Test	1.84	1.02
Period 3	Train	1.33	0.61
	Test	1.78	0.87
Period 4	Train	1.38	0.60
	Test	1.69	0.86

Table 4.1: Mean and standard deviation of MAE per period

Consequently, aberrant values were observed in 276 heaters, while 73 of them had been measured only for two heating periods, and thus their reliability was questioned due to possible insufficiency of data in one or both periods. Some of these models were trained or/and tested for less than 60 days per period, which may be

considered insufficient for RF, whereas few of them were trained and tested for several days and might have performed better in a simpler model, such as LR (Agrawal 2022). It is noteworthy that 7 heaters expounded significant deviations in two pairs of periods, but these were heaters measured for a shorter time in two or three of the heating periods as well.

Some of these heaters were examined individually to inspect whether the error distribution of their train and test period was noticeably different, but the results demonstrated unusual behavior mostly in cases of few data points. For instance, heater 32567, as shown in figure 4.4, presents discrepancies in more than one pair of periods, but its set sizes per period were quite unbalanced. As evidenced by table 4.2 and 4.3, periods 1 and 4 had the fewest data and the greatest difference between the MAE of the two sets, regardless of their use as train or test set.

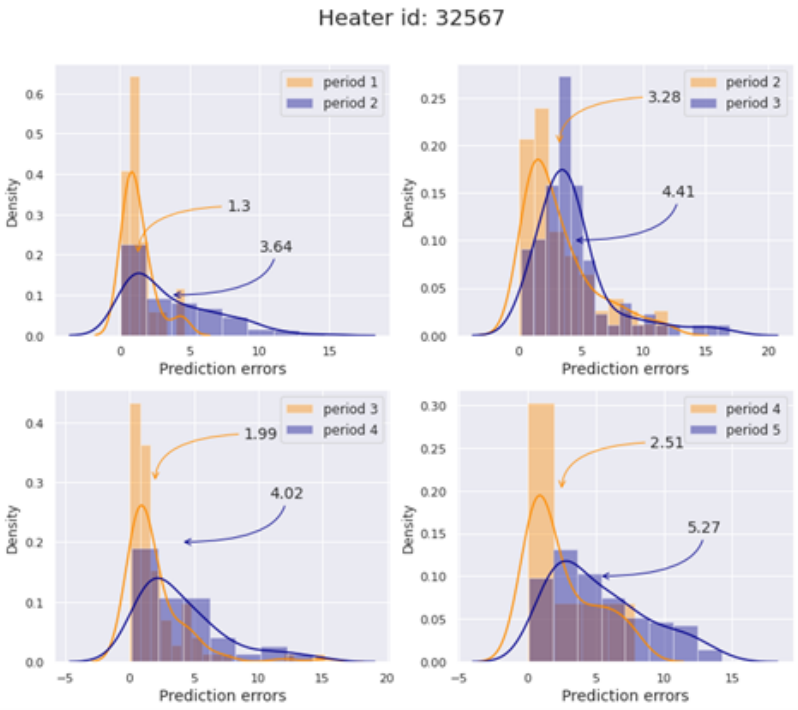


Figure 4.4: Heater 32567 - prediction errors per period.

Period	No. of days measured
1	26
2	129
3	90
4	37
5	120

Table 4.2: No. of days measured per period - heater 32567.

Train period	Test period	MAE train	MAE test
1	2	1.30	3.65
2	3	3.28	4.41
3	4	1.99	4.02
4	5	2.51	5.27

Table 4.3: Test and train MAE per period - heater 32567.

Given that specific heaters were already suspected by the results of LR (Agrawal 2022) and SVR (Münten 2022) of possible improvement in insulation, the density of their prediction errors was plotted to identify whether the initial assumption, regarding the means difference, was valid. Therefore, the residuals' distribution of the heaters 8180 and 27729 are shown in Figures 4.5 and 4.6, respectively. As stated previously, heater 8180 displayed a change in gas use during the third period, while heater 27729 showed analogous deviation during the fourth period. The same conclusion could be drawn from the residuals' distribution of these two heaters for each period, while the difference between the means of the train and test sets was quite evident and statistically significant.

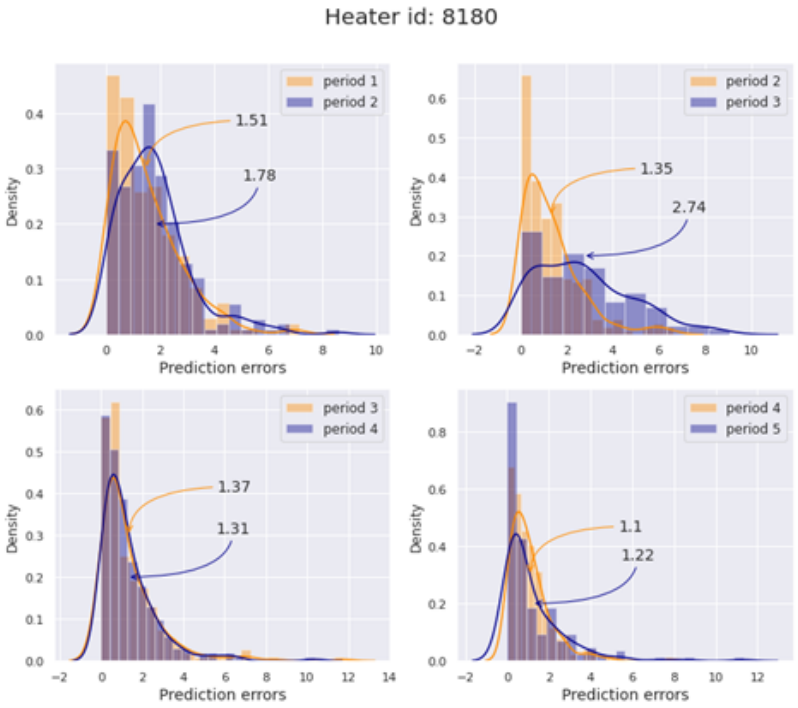


Figure 4.5: Heater 8180 - prediction errors per period.

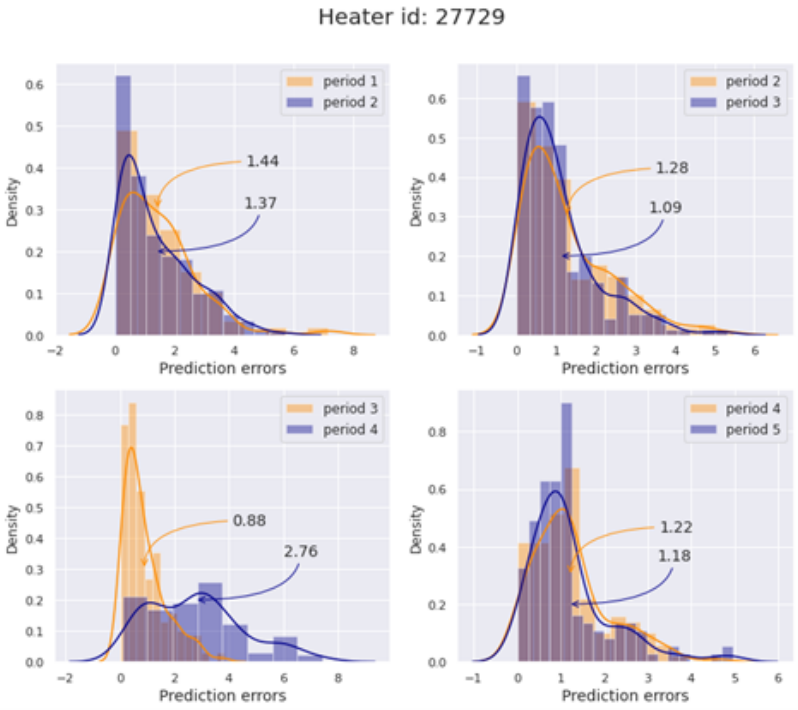


Figure 4.6: Heater 27729 - prediction errors per period.

Alternatively, after training the model with data for every period, models for every month, period, and heater were built to explore whether an improvement in insulation could be identified earlier. As expected, the specific analysis was more time-consuming than the preceding experiment and hence was only applied to 6000 heaters from which only 4482 complied with the essential number of data points. The approximate time of hyperparameter tuning and training for every heater was 2,5 minutes. The analysis of the previous case indicated that RF did not perform well for small data sets and biased this investigation.

Furthermore, t-tests were performed for every pair of prediction errors, since their sample size was equal to 31 days or shorter, and it was observed that 4198 of these heaters had statistically different measurements per month and period. The MAE was also inspected and resulted in outrageous test scores. As it was inconvenient to inspect the score of each month and each period individually, figure 4.7 shows the MAE of every period, and validates that these models performed generally worse than the period-wise models. Thus, the present experiment was not completed due to these early signs that the results were inaccurate.

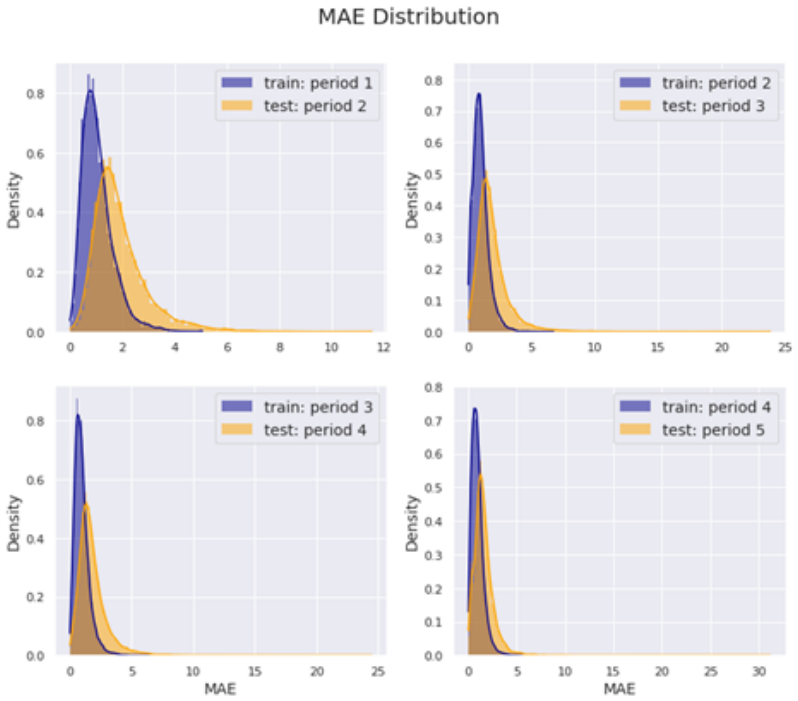


Figure 4.7: MAE Distribution per period of models trained with monthly data.

5 | Conclusion and Discussion

The Random Forest Regression was analyzed, as a part of a widened research, to support Intergas' to act on the climate change by reducing the energy consumption. Even though the model is not linear and could not be directly compared to LR (Agrawal 2022) and SVR (Münten 2022), it was considered a supplementary study to discover possible relevance.

Random Forest was trained using several different numbers of data points, given the availability of them by Intergas, and proved to be unsuitable for the current analysis. Due to its nature, the model does not reveal much information about its internal operations, as it cannot be interpreted in terms of coefficients and intercept. Moreover, it was noticed that the RF model produced larger MAE for higher parameter values, and it was extremely sensitive to the sample size of the train and test set.

Most of the pairs of train and test sets appeared to be statistically different and produced abnormal MAE scores in cases that the data were not enough for RF to become a stable predictor. Because of this complication, there was no proof that RF could be used to discover the change in insulation using fewer data points and this limitation evinced that the current research was better approached by the linear models.

During the research, the fast detection of the change in daily gas use was of crucial importance, and thus, models were fit into data of every month, period, and heater to compare the gas use of the months between two consecutive periods. The results showed that the amount of information is inadequate for RF, and besides the quantity of the data had been an issue of the specific model for longer periods too.

Further research in the specific domain could be amended by including more factors that might influence the gas consumption, such as the weather conditions and personalized behavior of the users. The gas use and temperature difference were positively correlated but obviously the gas use could not be fully explained by the insulation. In such circumstances, RF may produce better results and exhibit its strength, as it is identifiable for its ability to handle complex datasets.

5.1 Comparison of Models

In a comparison of the three different approaches and models in the various works, the following result could be achieved. It is clear that Varoon Sushil Agrawal's approach of a linear regression method is very similar to Moritz Muenten's support vector regression model. This indicates that both models show the linear relationship between the average temperature difference and the added gas consumption per day. Also in the results, despite different approaches to filtering and distribution, there is a large overlap in the final selection of heaters with valid decrease and potential households where an insulation could be a reason for that. Maria Fakou's approach of detecting significant changes in the heaters using a non-linear model such as a random forest shows that this approach was able to detect the various heaters that come into question, but the breadth of the results due to other error rates is so high that one cannot obtain a valid result.

References

Agrawal, V. S. (2022), 'Doing more with less - 1', *Utrecht University* .

Brownlee, J. (2018), 'A Gentle Introduction to k-fold Cross-Validation', https://machinelearningmastery.com/k-fold-cross-validation/. [Online; accessed 23-May-2018].

Consumption of energy (2016), https://ec-europa-eu.proxy.library.uu.nl/eurostat/statistics-explained/index.php?title=Consumption_of_energy. [Online; accessed 29-June-2022].

Delft CE, Hinicio, I. I. E. C. D.-G. f. E. (2015), 'Financing the energy renovation of buildings with cohesion policy funding : technical guidance: final report', *Publications Office of the European Union* .

Faidra Filippidou, Nico Nieboer, H. V. (2018), 'Effectiveness of energy renovations: a reassessment based on actual consumption savings', *Effectiveness of energy renovations: a reassessment based on actual consumption savings* .

Haris Lulic, Adnan Civic, M. P. A. O.-E. D. (2013), 'Optimization of thermal insulation and regression analysis of fuel consumption', *24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013* .

Hayes, A. (2022), 'Empirical Rule', https://www.investopedia.com/terms/e/empirical-rule.asp. [Online; accessed 05-March-2022].

Koehrsen, W. (2018), 'Hyperparameter Tuning the Random Forest in Python', https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74#:~:text=A%20Brief%20Explanation%20of%20Hyperparamete%20Tuning&text=(The%20parameters%20of%20a%20random,be%20optimal%20for%20a%20problem. [Online; accessed 10-Jan-2018].

Mean absolute error (2022), https://en.wikipedia.org/wiki/Mean_absolute_error. [Online; accessed 13-May-2022].

Mean squared error (2022), https://en.wikipedia.org/wiki/Mean_squared_error. [Online; accessed 14-June-2022].

Münten, M. (2022), 'Doing more with less - 1', *Utrecht University* .

Schonlau, M., Z. R. Y. (2020), 'The random forest algorithm for statistical learning', *The Stata Journal, 20(1)* p. 3–29.

T-test (2022), https://en.wikipedia.org/wiki/Test_statistic. [Online; accessed 29-May-2022].

Z-test (2022), https://en.wikipedia.org/wiki/Z-test. [Online; accessed 18-April-2022].

Appendix

A | Full data exploration results

```
1
2  # Ig_gasuse_hourly filtering
3  gasuse_df = gasuse_df.filter((gasuse_df.oppervlakteverblijfsobject >= 40) & (
       gasuse_df.oppervlakteverblijfsobject <= 400) & (gasuse_df.t_set <= 26) & (
       gasuse_df.t_act <= 30) & (gasuse_df.t_act >= 10))
4
5
6  # filter out summer months from Ig_gasuse_hourly and od_knmi_hourly_wijken_v2
7  gasuse_no_summer = gasuse_df.filter((gasuse_df.month > 8) | (gasuse_df.month <
       5)).drop('month')
8  knmi_no_summer = knmi_hourly_df.filter((knmi_hourly_df.month > 8) | (
       knmi_hourly_df.month < 5)).drop('month')
9
10
11 # remove missing values from  ig-heater-info-nl-2 and house_prop
12 house_prop_df = house_prop_df.na.drop()
13 heater_info = heater_info.na.drop(subset=['pandbouwjaar','wijk']).select('
       HEATER_ID', 'pandbouwjaar', 'wijk')
14
15
16 # join the datasets
17 gasuse = gasuse_no_summer.join(heater_info, gasuse_no_summer.heater_id ==
       heater_info.HEATER_ID, "inner").drop(heater_info.HEATER_ID)
18 gasuse_with_knmi = gasuse.join(knmi_no_summer, ['Wijk', 'TimeKey'], "inner")
19 df_joined = gasuse_with_knmi.join(house_prop_df, gasuse_with_knmi.heater_id ==
       house_prop_df.HEATER_ID, "inner").drop(house_prop_df.HEATER_ID)
20
21 # removes heaters that contain multiple different records for the same date
22 duplicate_id = df_joined.groupby(['heater_id', 'TimeKey']).count() \
23     .where('count > 1').select('heater_id').distinct()
24 duplicate_id = [row[0] for row in duplicate_id.select('heater_id').collect()]
25 df = df_joined.filter(~df_joined.heater_id.isin(duplicate_id))
```

Figure A.1: data_cleaning.ipynb

B | Annotated scripts of analyses and method settings

```
results = []
for h in heaters:
    sample_data = data[data.heater_id == h]
    for period in range(1, 5):
        data_train = sample_data[sample_data.period == period]
        data_test = sample_data[sample_data.period == period+1]
        if(data_train.shape[0] >= 6) & (data_test.shape[0]>= 6):
            # assign train and test sets
            x_train = data_train.avg_t_diff.array.reshape(-1,1)
            y_train = data_train['sum_gas']
            x_test = data_test.avg_t_diff.array.reshape(-1,1)
            y_test = data_test['sum_gas']
            # Create the parameter grid based on the results of random search
            param_grid = {'max_depth': [1, 2, 3],
                'n_estimators': [1, 2, 3, 4, 5, 6]}
            # Create a based model
            rf = RandomForestRegressor()
            # Instantiate the grid search model
            grid_search = GridSearchCV(estimator = rf, param_grid = param_grid,
                cv = 3, n_jobs = -1, verbose = 2)
            grid_search.fit(x_train, y_train)
            # fit RF with best parameters
            rf_2 = RandomForestRegressor(max_depth=grid_search.best_params_['
               max_depth'], n_estimators=grid_search.best_params_['n_estimators
               '] )

            rf_2.fit(x_train, y_train)
            pred_train = rf_2.predict(x_train)
            pred_test = rf_2.predict(x_test)
            # compute prediction errors
            pred_errors_train = (y_train - pred_train).tolist()
            pred_errors_test = (y_test - pred_test).tolist()

            # create dictionary of complete information for every model
            row = {'heater_id': h, 'period': period, 'y_train': y_train.tolist
                (), 'prediction_train': pred_train.tolist(), 'y_test': y_test.
               tolist(), 'prediction_test': pred_test.tolist(), 'max_depth':
               grid_search.best_params_['max_depth'], 'n_trees': grid_search.
               best_params_['n_estimators'], 'pred_errors_train':
               pred_errors_train, 'pred_errors_test': pred_errors_test}
            results.append(row)
```

Figure B.1: createDatasetOfResuts-rf.ipynb

```
from sklearn.metrics import mean_absolute_error,
from math import sqrt

mae_tr,  mae_test = [], []

for index, row in df.iterrows():
    mae_tr.append(mean_absolute_error(row['y_train'], row['prediction_train']))

    mae_test.append(mean_absolute_error(row['y_test'], row['prediction_test']))

df['mae_train'] = mae_tr
df['mae_test'] = mae_test
```

Figure B.2: calculate_mae.ipynb

```
mean_mae_train_p = []
mean_mae_test_p = []
sd_mae_train_p = []
sd_mae_test_p = []

for i in df.period.unique():
    mean_mae_train_p.append(df[df.period==i]['mae_train'].mean())
    mean_mae_test_p.append(df[df.period==i]['mae_test'].mean())

    sd_mae_train_p.append(df[df.period==i]['mae_train'].std())
    sd_mae_test_p.append(df[df.period==i]['mae_test'].std())

three_s_train = np.add(mean_mae_train_p, [element * 3 for element in
    sd_mae_train_p])
three_s_test = np.add(mean_mae_test_p, [element * 3 for element in
    sd_mae_test_p])

heaters = []
for i, n in enumerate(df.period.unique()):
    temp = df[(df.period==n) & (df.mae_test>three_s_test[i]) & (df.mae_train<=
        three_s_train[i])]
    heaters.append(temp.heater_id.unique())

flatten_list = [element for sublist in heaters for element in sublist]
```

Figure B.3: mae_outliers.ipynb

```
from scipy import stats
from random import sample
from statsmodels.stats.weightstats import ztest as ztest

stat = []
p_values = []

stat_z = []
p_values_z = []

for index, row in df.iterrows():
    if (len(row['pred_errors_train'])>50) & (len(row['pred_errors_test'])>50):
        s_z, p_z = ztest(row['pred_errors_train'], row['pred_errors_test'])

        s, p = np.nan, np.nan
    elif (len(row['pred_errors_train'])<=50) | (len(row['pred_errors_test'])
        <=50) :
        s, p = stats.ttest_ind(row['pred_errors_train'], row['pred_errors_test'
            ], equal_var=False)

        s_z, p_z = np.nan, np.nan

    stat.append(s)
    p_values.append(p)

    stat_z.append(s_z)
    p_values_z.append(p_z)


df['stat'] = stat
df['p_value'] = p_values


df['stat_z'] = stat_z
df['p_values_z'] = p_values_z
```

Figure B.4: stat_tests.ipynb

C | Full analysis results

	0	1	2	3
heater_id	30497	30497	30497	30497
period	1	2	3	4
y_train	[4.96, 9.5, 2.0...	[8.52, 7.79, 1.61...	[16.6, 0.77, 15.46...	[3.25, 18.02, 13.01...
prediction_train	[4.72, 8.25, 4.72...	[8.4, 4.6, 4.18...	[18.09, 1.46, 16.02...	[7.86, 12.44, 14.29...
y_test	[8.52, 7.8, 1.61...	[16.6, 0.77, 15.46...	[3.25, 18.02, 13.01...	[10.86, 1.19, 5.03...
prediction_test	[4.72, 4.72, 4.72...	[16.42, 1.29, 16.41...	[11.4, 11.87, 11.87...	[7.86, 1.99, 6.78...
max_depth	1	3	3	2
n_trees	4	3	6	3
pred_errors_train	[0.24, 1.25, 2.71...	[0.13, 3.2, 2.57...	[1.5, 0.69, 0.57...	[4.62, 5.58, 1.27...
pred_errors_test	[3.8, 3.07, 3.11...	[0.17, 0.52, 0.96...	[8.15, 6.15, 1.14339...	[3, 0.81, 1.75...
mae_train	2.42	1.82	1.73	2.29
mae_test	3.13	1.93	2.66	2.238157
stat	NaN	NaN	NaN	NaN
p_value	NaN	NaN	NaN	NaN
stat_z	-2.52	-0.65	-5.16	0.27
p_values_z	1.183918e-02	5.152963e-01	2.446828e-07	7.876512e-01

Table C.1: Transposed table of results (first 4 rows)