Doing More with Less - 1

Master Thesis

Moritz Münten

Utrecht, July 1, 2022

Information Page Graduation Report

Utrecht University
Heidelberglaan 8, 3584 CS Utrecht, Netherlands

Master Thesis

Name of student:	Moritz Münten
Student number:	1149423
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: 5466

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, Moritz Münten (full name),

have read the agreement and will stick to it.

Date: 28.04.2022

Signature: Moritz Münten

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 each make a conclusion about the study, answer the research question with their approach, 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:	Moritz Münten
Student Number:	1149423
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.

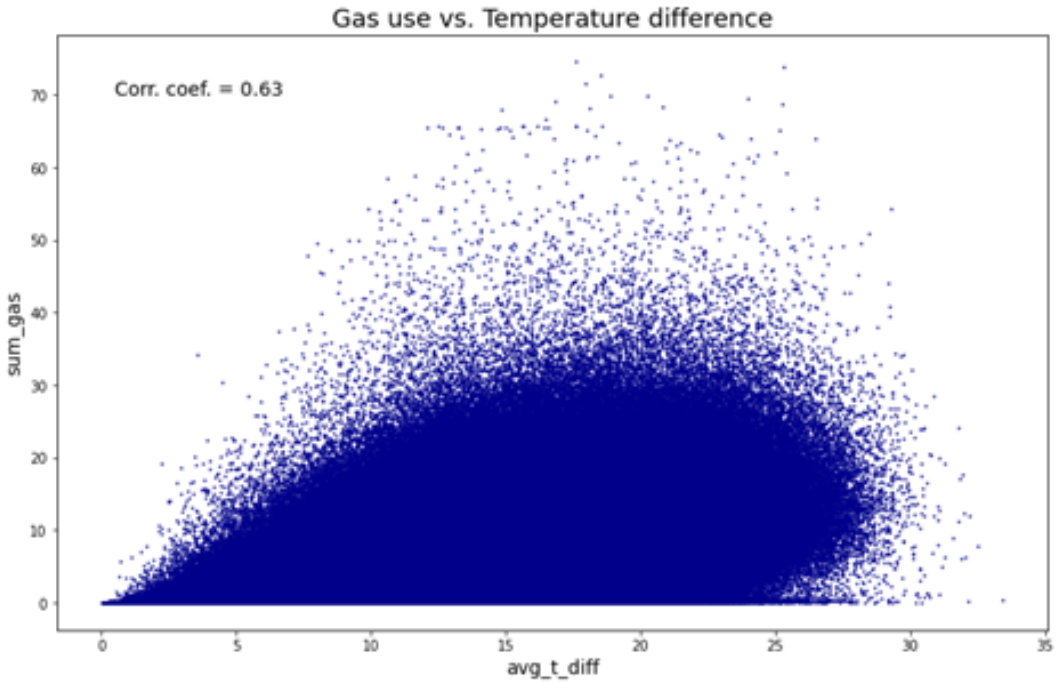

Figure 2.1: Gas use vs. temperature difference.

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.

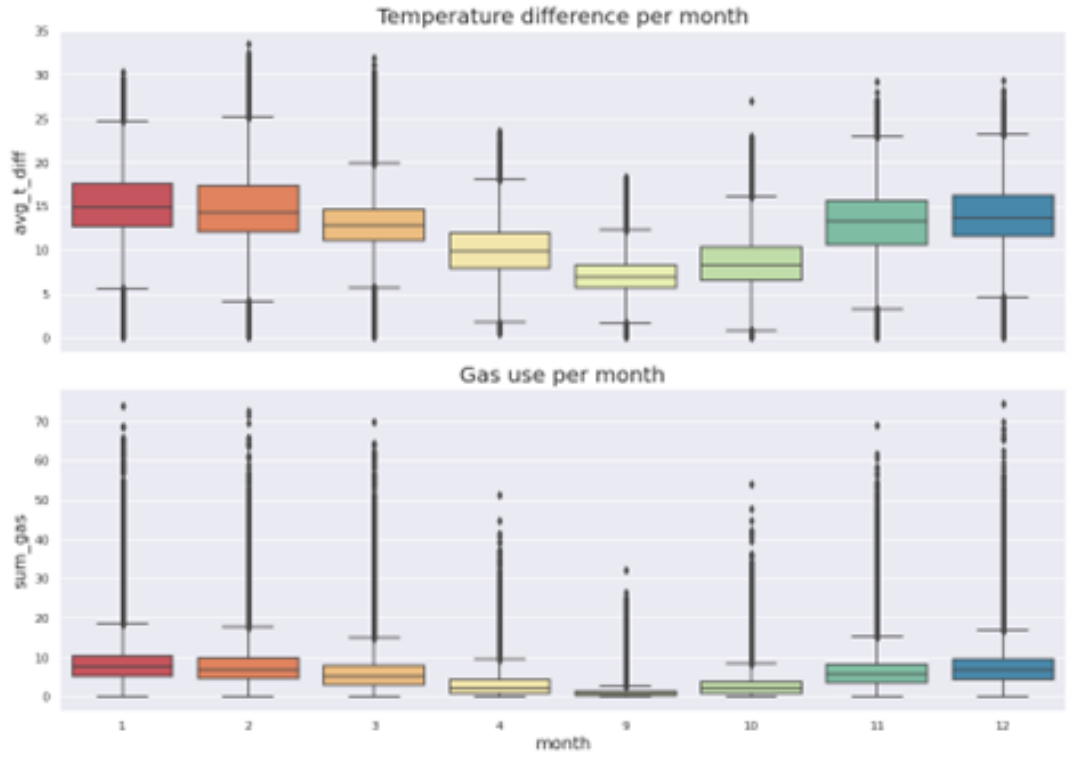

Figure 2.2: Temp. diff. and gas use per month.

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.

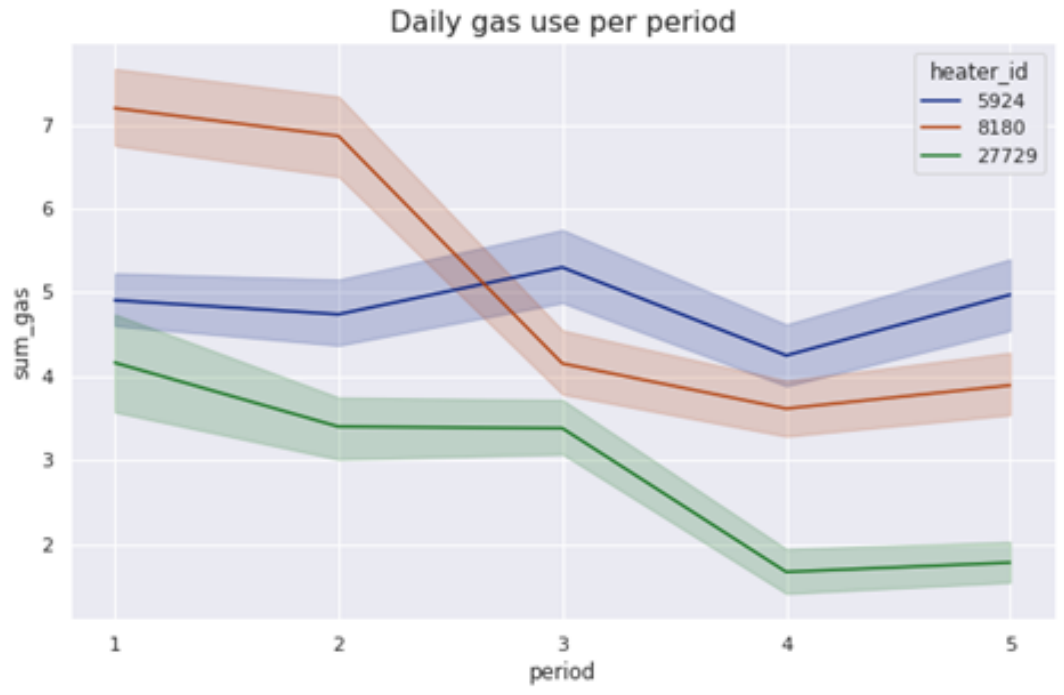

Figure 2.3: Daily gas use per period.

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

Now that we have assigned the two parameters sum_gas_use and avg_t_diff to the periods months and days. In this chapter, we will now perform a linear regression algorithm called Support vector regression to determine the coefficients of the two values for each period. To detect a decrease in slopes for one heater, all available households (heater ids) are checked. To have a deep dive into the analysis results, multiple individual heaters are analyzed and filtered to find heaters where a newly added insulation could be the cause of the valid decrease.

3.1 Translation of the research question to a data science question

To answer the research question, one must first calculate the individual slopes of the individual households for the individual heating periods. Then we have to investigate how far the slopes from the different heating periods differ. Finally, one must investigate how early one can detect such a change and what significance a decrease in the slope can be possibly attributed to a newly added insulation of the house.

3.2 Motivated selection of method for analysis

A support vector machine SVM is a flexible machine learning model that can handle linear and non-linear classification tasks, regression and outlier detection and was first mentioned in 1995 by Wladimir Naumowitsch Wapnik (Vapnik 1995). SVM are actually well suited for the classification of complex data sets of small or medium size, however, an inversion of this algorithm are used for regression. For linear and non-linear regression, the SVM algorithm tries to place as many data points as possible on the road (epsilon width) and minimize boundary violations, i.e. data off this road. The width of the regression boundary is controlled via the hyperparameter Epsilon. Hyperparameter decides on the width of the margin on which the road is drawn. Damage from additional training data points within this margin does not affect the prediction of the model, so this model is called epsilon intensive (Geron 2018).

3.3 Motivated settings for selected method

As explained, for support vector regression, you first have to determine the appropriate epsilon (margin-width) value. The processed data of the Data wrangling was adjusted for this purpose. Later, the regression per period is to be calculated and predicted for a single household. In order to have comprehensive test data for the individual periods, all stimuli with less than 630 data points were first removed. Because a period reflects approx. 210 data points from 30 days of a month.

Now the time frame necessary to make valid statements has been determined. For this purpose, a linear regression was first used to obtain an initial overview of the percentage differences between the various slopes of the respective months. It became clear that the variation of the different months is too high that no reasonable conclusions could be drawn about a reliable increase or decrease in gas consumption. Accordingly

periodic slopes were calculated instead of monthly so that one could recognize at a better statement about the increase or decrease. Moreover, with this approach it became clear that one period is not sufficient to recognize whether it is a valid decrease, as one needs a comparative value from a previous or following period. Finally, two heating periods are still not sufficient to speak with certainty of a decrease in gas consumption, since a decrease in gas consumption can be determined between two periods, but this is only valid if it remains at a constant low level in the following heating period. Thus, one can also exclude the possibility that a decline between two periods can be attributed to insufficient or incorrect data because the change stays constant in the following period.

The result is that you can only confirm a significant decrease in gas consumption with at least three periods of data available, because there one can only calculate two differences between the respective individual periods and thus recognize whether in the case of a first decrease between the first two periods this remains constant and is thus valid. This number of data points was determined in order to find the correct hyper-parameters for the SVR model. The filtered dataset has the following properties(see Table 3.1: Filtered dataset for model tuning summary).

	heater_counts
count	4829.000000
mean	824.554152
std	141.184528
min	631.000000
25%	692.000000
50%	810.000000
75%	927.000000
max	1103.000000

Table 3.1: Filtered dataset for model tuning summary

To select a representative heater for our SVR tests, a heater was selected that had approximately the same number of data points as the average of 824 of the dataset. The heater 27729 with 846 entries was selected for this.

Therefore, we try to find the best parameters with evaluation tests with the data of the heater 27729. All data points were first allocated to the individual periods, resulting in five different data frames (see appendix B.1: createDataSetsPerHeater_svm2.ipynb[9]). Next, a training set and a test set were split for each individual period, with the distribution being 80% of the data belonging to the training set and 20% of the data to the test set (see appendix B.2: splitTrainAndTestSetsFromDataset_svm2.ipynb[10]), which is a standard division for this training method (Geron 2018). Then copies of the individual data were made and the individual values for the sum of the gas consumption and average temperature difference were filtered out (see appendix B.3: filterValuesFromPreparedTestandTrainSet_svm2.ipynb[11]) in order to train them with the linear SVR model. The different models were tested with different values for Epsilon (see appendix B.4: trainModelForEachPeriodWithEpsilonValue_svm2.ipynb[12]), with the ranch ranging from 0.2 to 2 for Y (out of this range, the error rate always became higher). In each run, the predictions of the model were calculated with the test data set of the respective period to obtain the linear mean square error. The root-mean-square error (RMSE) was then calculated, which can make a statement about the accuracy of the model of the predictions (see appendix B.5: rmseCalculationForPeriod1_svm2.ipynb[13]). By adjusting the epsilon value in the prescribed range, an attempt was made to determine the smallest possible RMSE for each period. The result of this analysis showed that the best y-value for the given data is 0.5. Below in figure 3.1: SVR Predictions for Periods is the exact hit rate of the model using the training and test data, as well as a variation plot over the RMSE for each period.

However, one can see a clear tendency which reflects the linear relationship between the average temperature difference and the gas consumption. The drawn support vector regression is therefore a valid approach for the prediction and analysis of the consumption of the individual heaters.

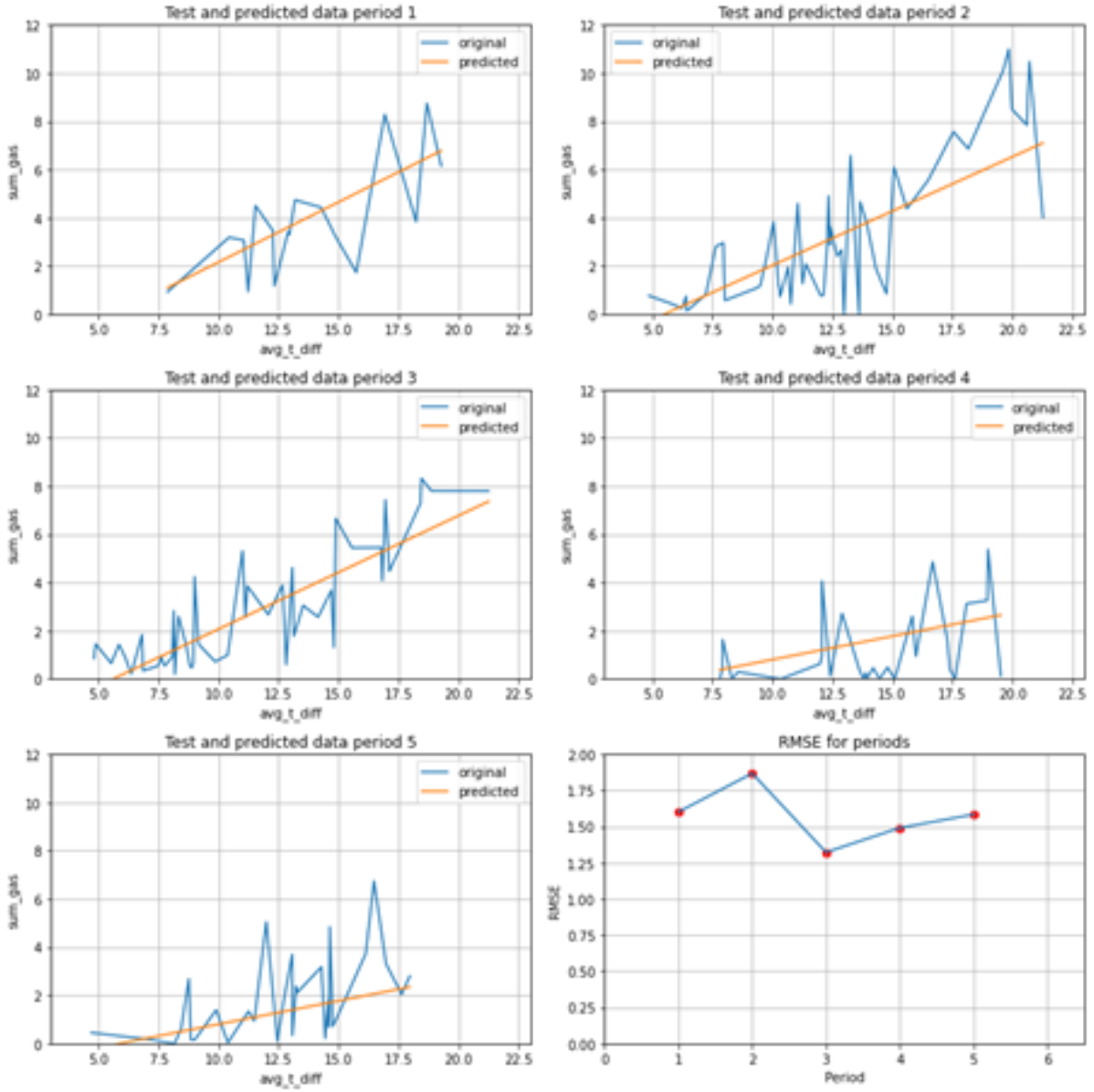

Figure 3.1: SVR Period Predictions for heater 27729

4 | Results

Now that we have determined the optimal settings for our support vector regression, we will calculate the slopes for each period from each heater, where enough data is available. Hereby the intercept is ignored since this study is focused only on comparing slopes for gas use. We directly determine the difference in slopes between the individual periods as the percentage difference between them and filter out coefficient values that are below 0.1 since duo to have better results (see appendix B.6: calculatingSlopesForEachHeaterPeriod-svm_final.ipynb[6], B.7: calculatingSlopesForEachHeaterPeriod-svm_final.ipynb[7], B.8: calculatingSlopesForEachHeaterPeriod-svm_final.ipynb[8]). Below in Table 4.1: Preview Evaluated Data Part 1 and Table 4.2: Preview Evaluated Data Part 2 are a preview of the resulting dataset.

	_c0	heaterid	slope_p1	slope_p2	slope_p3	slope_p4	slope_p5
0	0	93059.0	NaN	NaN	0.503504	0.536394	0.465365
1	1	96265.0	NaN	NaN	0.365956	0.362242	0.397795
2	2	66595.0	NaN	0.561229	0.634732	0.599158	0.603931
3	3	54477.0	NaN	0.468828	0.509359	0.509249	0.525701
4	4	39755.0	NaN	0.742660	0.721017	0.669719	0.692452
...	...	...	...	...	...	...	...
12670	12670	17168.0	0.203675	NaN	NaN	NaN	NaN
12671	12671	20940.0	NaN	NaN	NaN	NaN	NaN
12672	12672	177291.0	NaN	NaN	NaN	NaN	0.194137
12673	12673	22093.0	NaN	0.116032	NaN	0.113248	NaN
12674	12674	72303.0	NaN	NaN	NaN	NaN	NaN

Table 4.1: Preview Evaluated Data Part 1

diff_1	pdiff_1	diff_2	pdiff_2	diff_3	pdiff_3	diff_4	pdiff_4
NaN	NaN	NaN	NaN	0.032890	6.532200	-0.071029	-13.241880
NaN	NaN	NaN	NaN	-0.003714	-1.014772	0.035553	9.814736
NaN	NaN	0.073503	13.096833	-0.035574	-5.604563	0.004773	0.796586
NaN	NaN	0.040531	8.645172	-0.000110	-0.021569	0.016452	3.230614
NaN	NaN	-0.021643	-2.914305	-0.051298	-7.114621	0.022733	3.394378
...	...	...	...	...	...	...	...
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 4.2: Preview Evaluated Data Part 2

As can be seen from the graph, there are multiple heaters who have no values for the respective individual slopes and consequently no values for the respective differences. This results either from a lack of data for

calculating the slope or from the fact that data is only available for a particular period. The following two graphs show how many null values are still in the elaborated data set (Table 4.3: Non-Null Count Data) that show coefficients to the results of the exploratory data analysis of the data available. The other two tables (Table 4.4: Information of dataset 1, and Table 4.5: Information of dataset 2) the exact properties of the final data frame. From this result, the following method is used to find the heaters that had a significant decrease in gas consumption in a period and thus be eligible for a possible added new insulation of the house.

Column	Non-Null Count	Dtype
_c0	12675	int32
heaterid	12675	float64
slope_p1	2869	float64
slope_p2	5874	float64
slope_p3	8663	float64
slope_p4	11186	float64
slope_p5	10250	float64
diff_1	2740	float64
pdiff_1	2740	float64
diff_2	5372	float64
pdiff_2	5372	float64
diff_3	7863	float64
pdiff_3	7863	float64
diff_4	10050	float64
pdiff_4	10050	float64

Table 4.3: Non-Null Count Data

	_c0	heaterid	slope_p1	slope_p2	slope_p3	slope_p4	slope_p5
mean	6337.000000	83647.559842	0.485091	0.602400	0.653769	0.645985	0.645804
std	3659.101666	51822.316012	0.221728	0.273383	0.290020	0.287830	0.278887
min	0.000000	2036.000000	0.101549	-0.195508	0.100130	0.100021	0.100221
25%	3168.500000	39992.000000	0.331790	0.420161	0.455548	0.451066	0.459645
50%	6337.000000	77199.000000	0.454709	0.568331	0.619824	0.612951	0.612709
75%	9505.500000	124790.000000	0.599577	0.744817	0.803980	0.795734	0.791104
max	12674.000000	204773.000000	1.970066	2.986652	2.867843	2.893415	4.630985

Table 4.4: Information of dataset 1

	diff_1	pdiff_1	diff_2	pdiff_2	diff_3	pdiff_3	diff_4	pdiff_4
mean	0.156091	41.266351	0.055368	17.707866	-0.002535	4.067250	-0.007086	2.635311
std	0.167089	57.126374	0.176085	58.416150	0.149239	47.006623	0.138739	31.168278
min	-0.748527	-85.517544	-1.365562	-244.142232	-1.106927	-91.156104	-1.501028	-89.592892
25%	0.065693	14.400163	-0.024653	-3.965965	-0.058172	-9.069329	-0.064698	-9.855553
50%	0.138561	29.390681	0.031865	5.380546	-0.003333	-0.506924	-0.006885	-1.184582
75%	0.227342	50.231546	0.111545	20.393284	0.051000	8.287111	0.051623	8.919707
max	2.462119	801.179597	1.538413	934.092805	1.356700	1200.815893	2.408475	583.081626

Table 4.5: Information of dataset 2

Now a density plot of the complete final dataset is computed to make a general assumption about the percentage changes that are significantly and the changes that are a normal variations in gas consumption. In this analysis, we make the assumption that all changes in the calculated percentage differences that lie within the 90% distribution are not significant changes and only a variation of gas consumption. Only the transitions of the upper and lower 5% of the distribution are classified as outliers and thus count as significant increases or decreases.

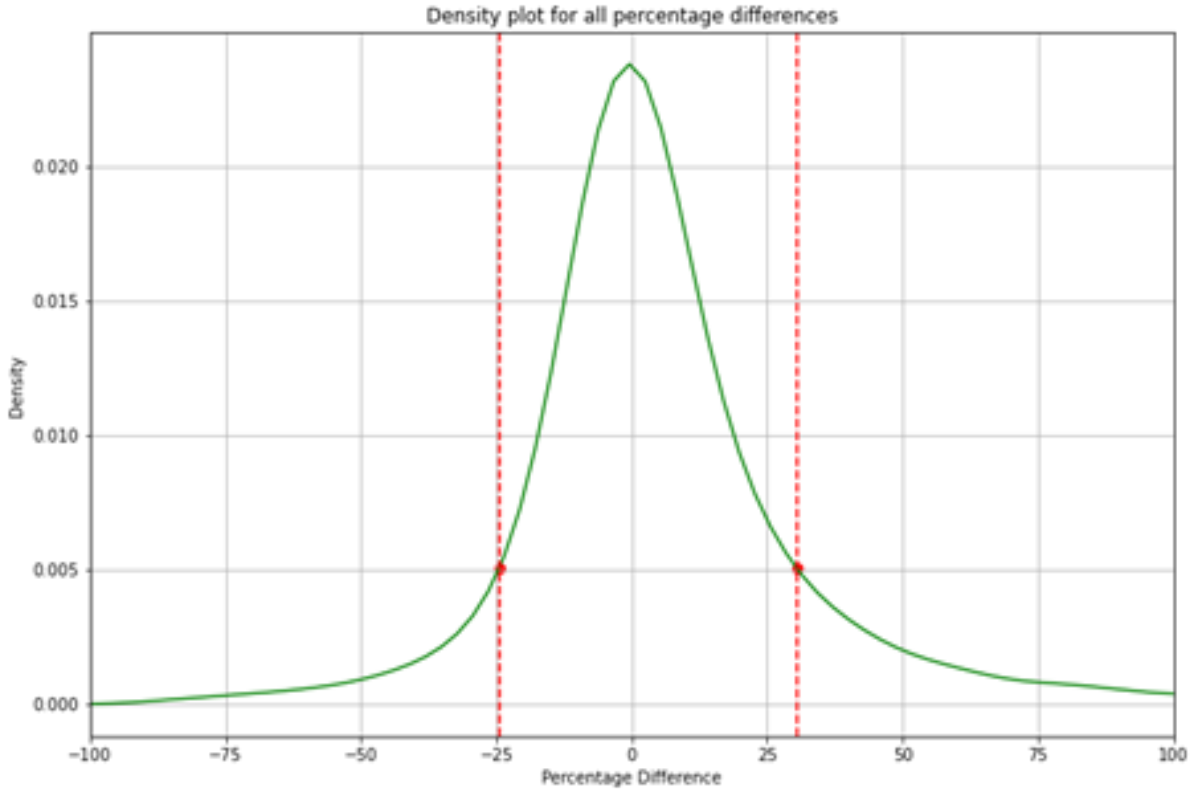

Figure 4.1: Density plot for all data

As can be seen from the distribution of increases and decreases, 90% of all values have a cycle between +30,5% and -24,5%. 10% of all values exceed or fall below this mark, which in this approach is defined as significant change in gas consumption.

4.1 Selected analysis results

Now, using the results obtained, a search function can be applied to the results to look for individual households that are eligible for a possible decrease in gas consumption. All slopes of each period that are negative and are due to erroneous data are first filtered out. Then one takes the distribution of the previous density plots to determine, on the basis of the latter, the percentage decrease of the rare as 5% appearance, which in this case is defined as a significant change. We make the assumption that in the following heating period the increase in gas consumption is not outside the 90% distribution, conclude that the data from the previous heating period was incorrect and is now being corrected. We add that no slopes of the periods have an increase above the 5%, which is due to a general unreliability of the data from that specific heater. In addition, only current heaters and those still used on the basis of the data are selected and the heaters that have no slot in the last or penultimate period are not taken into account, even if they may have had a decrease previously, they are no longer relevant. The last criterion of the filter is that a heater must have data for at least the last three periods and thus be more recent than one can determine with certainty a decrease, so that less than 3 slopes one cannot determine whether the decrease is constant or an ingredient of invalid data. An excerpt of this programmed filter function can be found in Appendix Figure C.1: filterHeatersWithValidDecrease-possibleHeaterDetection.ipynb[5]. A excerpt of the resulting dataset of this filter function as seen below (see appendix Table C.1: Complete final results preview part 1, and Table C.2: Complete final results preview part 2, for the full preview). The final dataset contains 202 individual households (heaters) that are eligible for a possible newly added insulation, as they show a valid decrease in gas consumption in a heating season with persistently lower gas consumption in the following periods.

heaterid	slope_p1	slope_p2	slope_p3	slope_p4	slope_p5
27729.0	0.485647	0.468805	0.480826	0.191385	0.212005
57721.0	NaN	0.414850	0.426396	0.275337	0.290224
8180.0	0.692461	0.717407	0.414945	0.468698	0.515741
45441.0	NaN	1.150599	0.583266	0.553134	0.634685
77589.0	NaN	NaN	0.801614	0.324421	0.359556

Table 4.6: Filter results preview

The following selection of the above 5 displayed heaters from the final dataset cover the different patterns, which are recognized by the percentage difference and a significant decrease by the filter function and give a detailed insight. To have a comparison to the 5 heaters first an insight into one heater is given that is not in the resulting dataset of the filter function, which shows a normal gas consumption with the variation we set (see appendix Table C.3: Complete data for heater 8941 part 1, and Table C.4: Complete data for heater 8941 part 2, for the full preview). For a better comparison of the slopes all lines start from (0,0) which is not the case for support vector regression lines.

heaterid	slope_p1	slope_p2	slope_p3	slope_p4	slope_p5
8941.0	0.433552	0.53911	0.551414	0.520482	0.55422

Table 4.7: Filter 8941 data

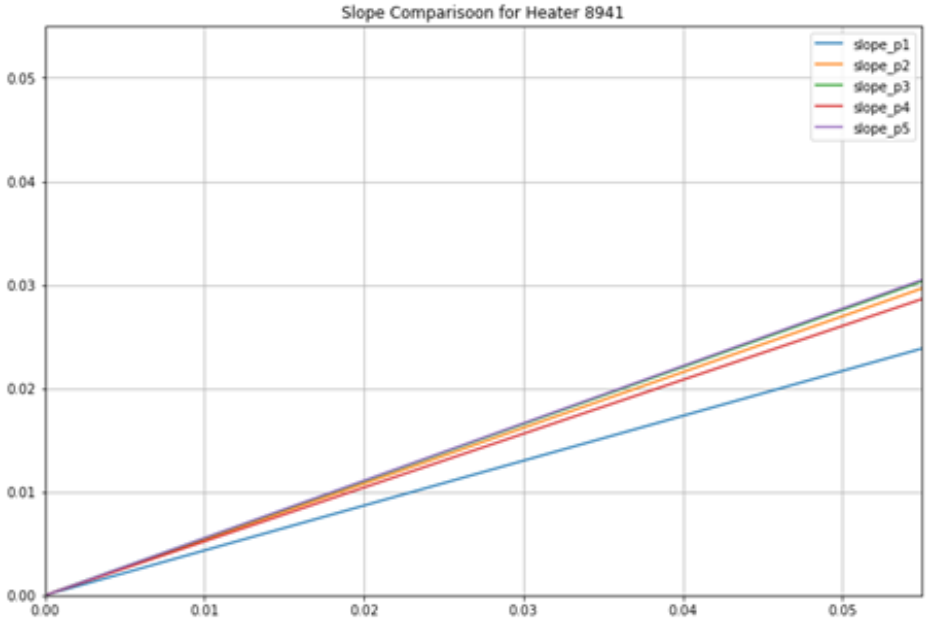

Figure 4.2: Slopes Heater 8941

If we now take a closer look at the percentage differences within the individual periods, we can see, as in Figure 4.3: Percentage Difference 8941, that the fluctuations in gas consumption vary within the 90% distribution between +30.5% and -24.5% and thus no significant decline can be detected.

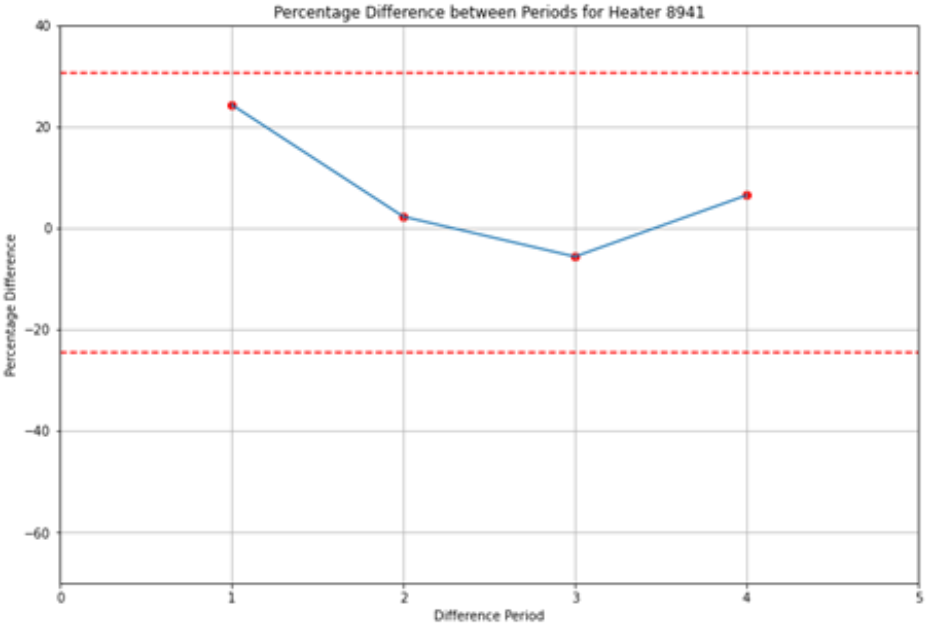

Figure 4.3: Percentage Difference 8941

If we now take a closer look at the slopes and percentage differences of the five heaters from our filtered dataset, you see the difference we are searching for. At heater 27729 it is clear that the first three heating periods are relatively the same and the gas consumption was accordingly similar. However, from heating period 4 onwards, the slope changes rapidly and shows a significant decrease. In addition, it remains constant lower in the last period.

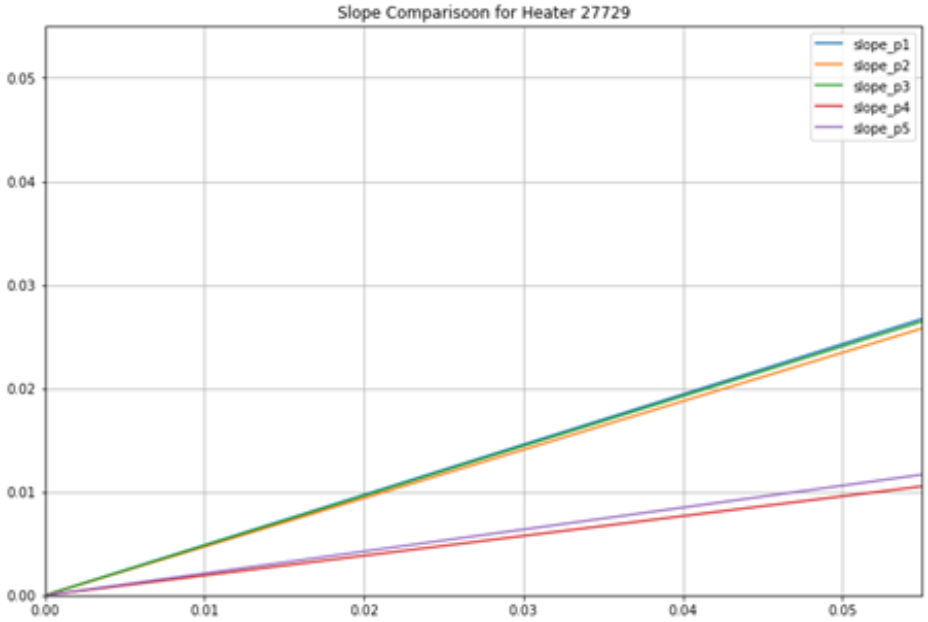

Figure 4.4: Slopes Heater 27729

Figure 4.3 below also reflects this in the percentage difference between the individual heating periods. To see that in periods 1 and 2 the slope is relatively constant to the gas consumption, however between periods 3 and 4 there is a decrease of -60% which is outside the normal change as shown on the density distribution before. This rapid decline, which exceeds our border of 5% density, which as already mentioned is below -24.5%, shows in comparison to the previous heater 8941 that a possible insulation has been added here.

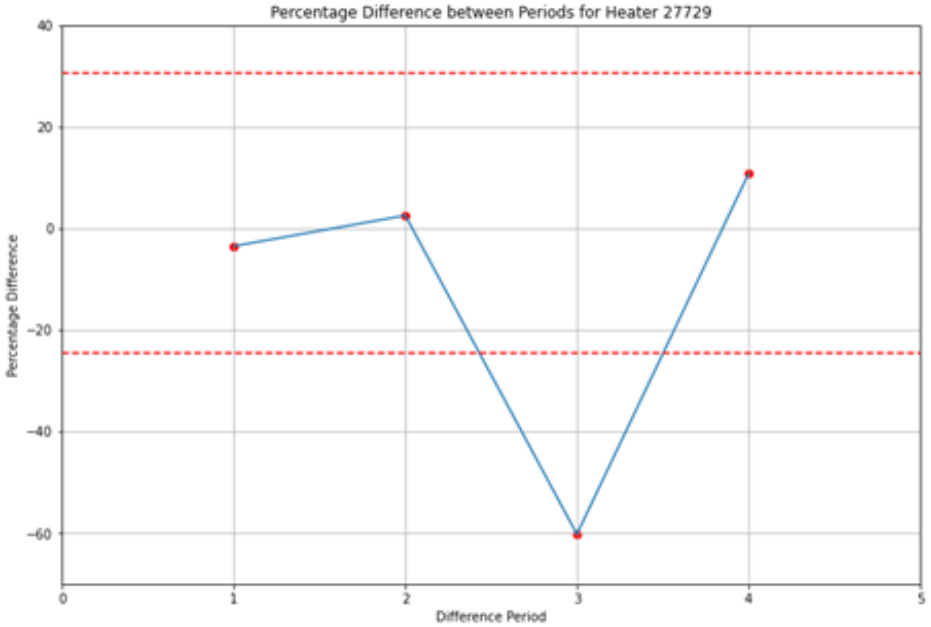

Figure 4.5: Percentage Difference 27729

A similar recognizable pattern can be found with heater 8180 from the results table. Figure 4.4 also shows very similar consumption in the first and second heating period. This changes in the third heating period, and the slow remains relatively constant at a lower level in heating periods 4 and 5.

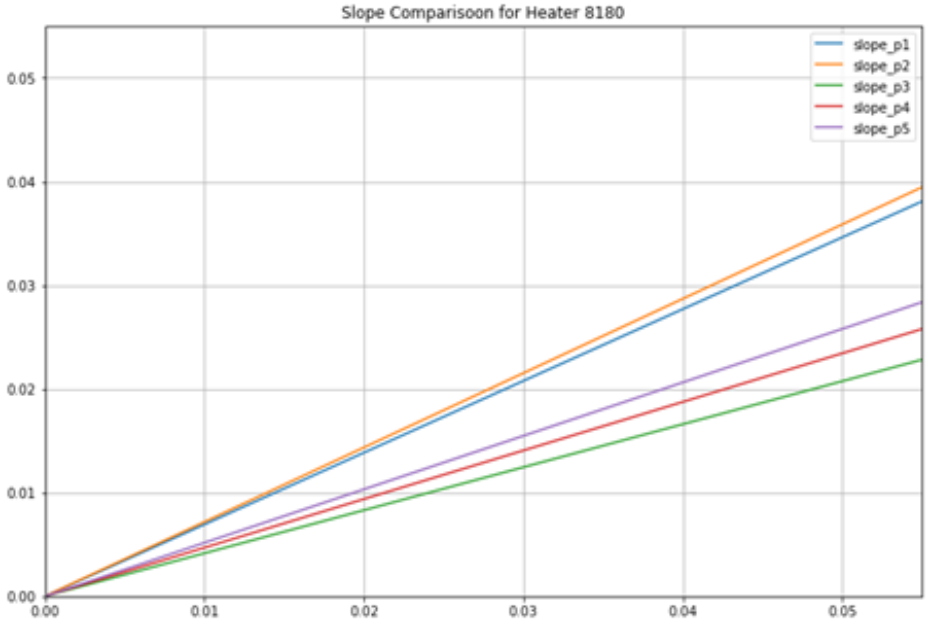

Figure 4.6: Slopes Heater 8180

The same pattern can already be seen in the previous result, which is shown in Figure 4.5, which shows that in the comparison between the first two heating periods the difference remains very small, but between the second and third period, as already mentioned, there is a decrease of -42%. Gas consumption rises again with 12% growth, but from the previous results we can see that an increase of less than 30.5% often indicates a normal change and is within limits. After, the slope and percentage change remains at a constant, lower level than before.

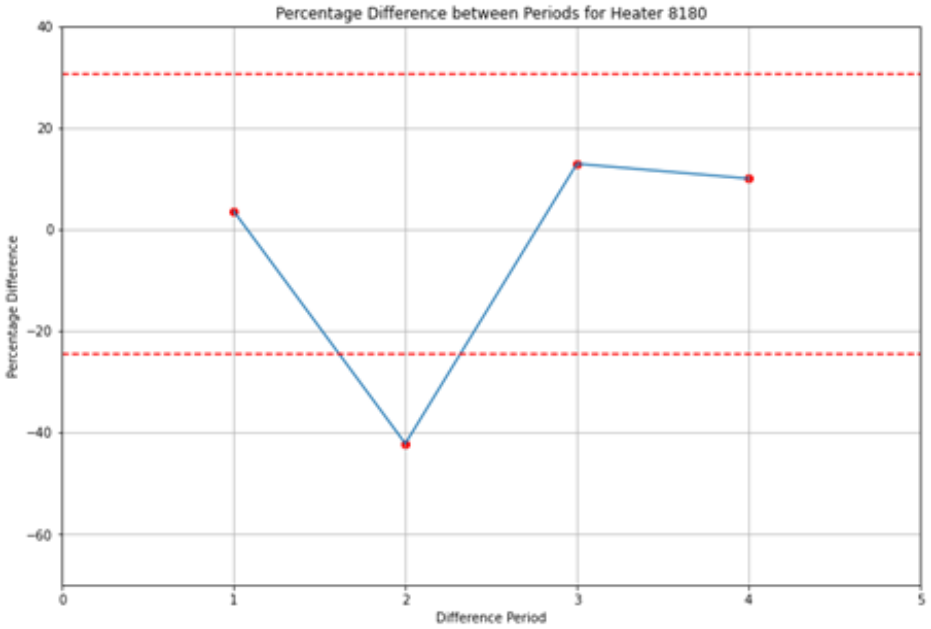

Figure 4.7: Percentage Difference 8180

In the further evaluation of the final data set, other examples of the bags are shown to show how different heaters are recognized by the filter and to subsequently show their pattern.

From the list we have heater 57721, which only has 4 periods of data available, but here the pattern can be seen that it has a significant drop after a constant consumption and then remains constant again.

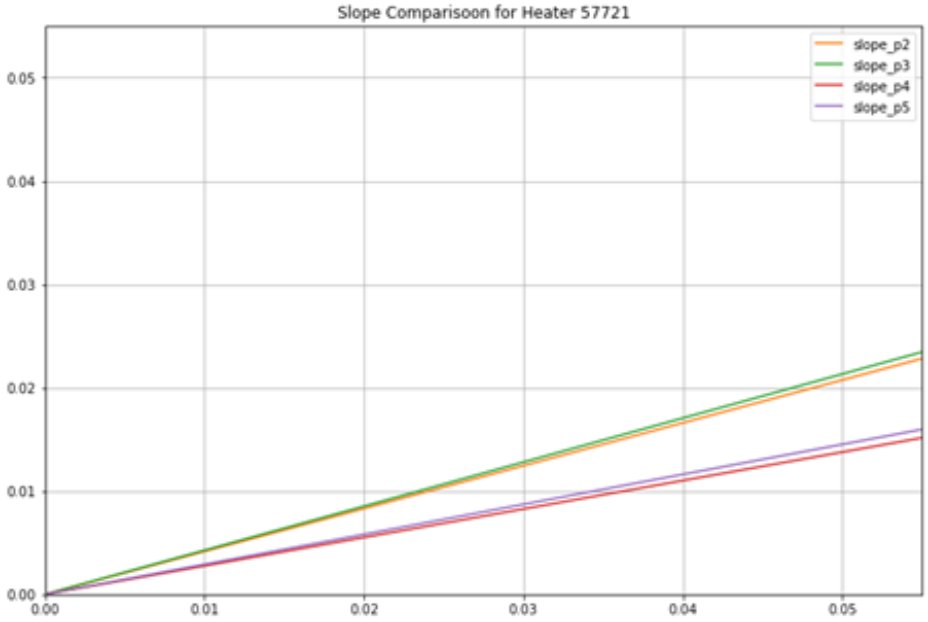

Figure 4.8: Slopes Heater 57721

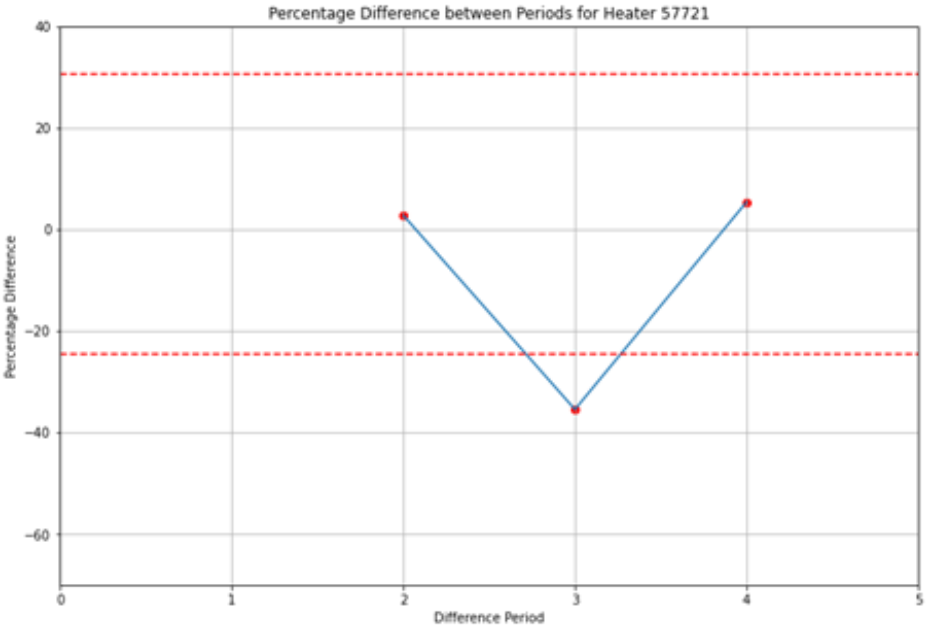

Figure 4.9: Percentage Difference 57721

The same pattern can also be seen in another sequence, for example the following heater 45441 that has also only four slopes, but here the decrease can already be seen between the first two heating period and then remains at a constant level in the normal variation.

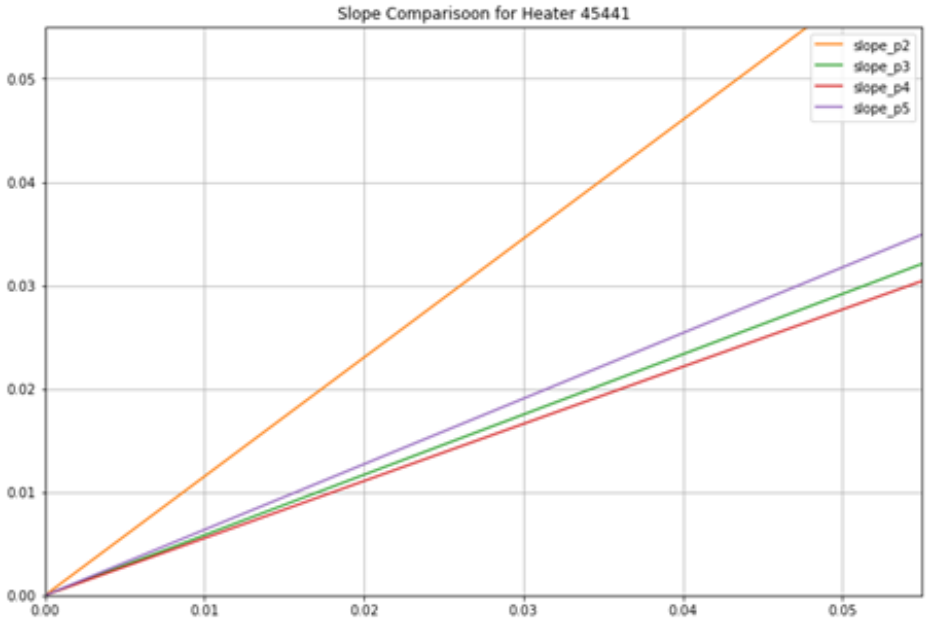

Figure 4.10: Slopes Heater 45441

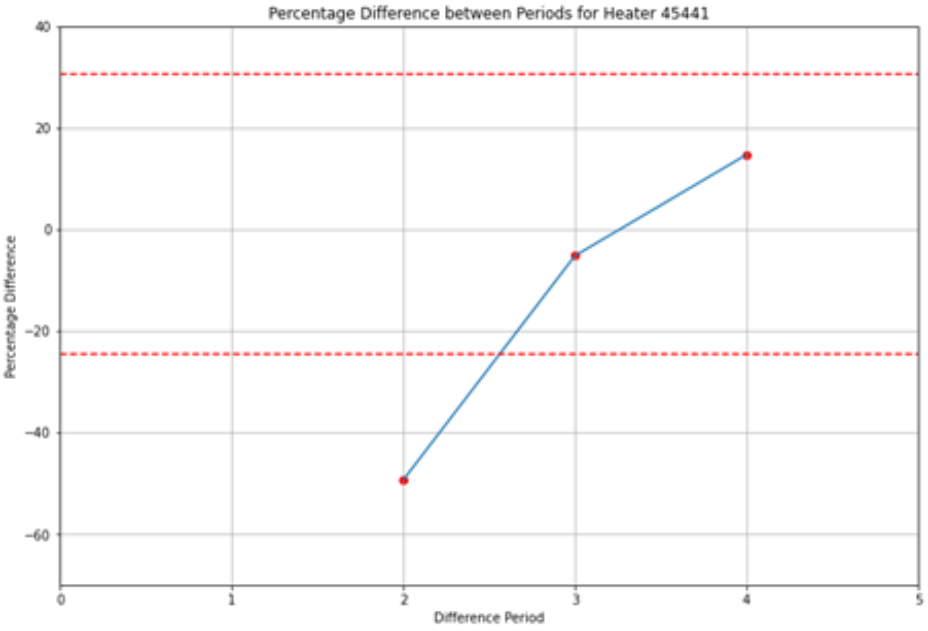

Figure 4.11: Percentage Difference 45441

Lastly, the diagrams for heater 77589 show the minimum of 3 heating periods that are need to recognize a valid decline. The second figure shows how the percentage decline between the 3rd and 4th period is outside the boundary and then remains at a constant level.

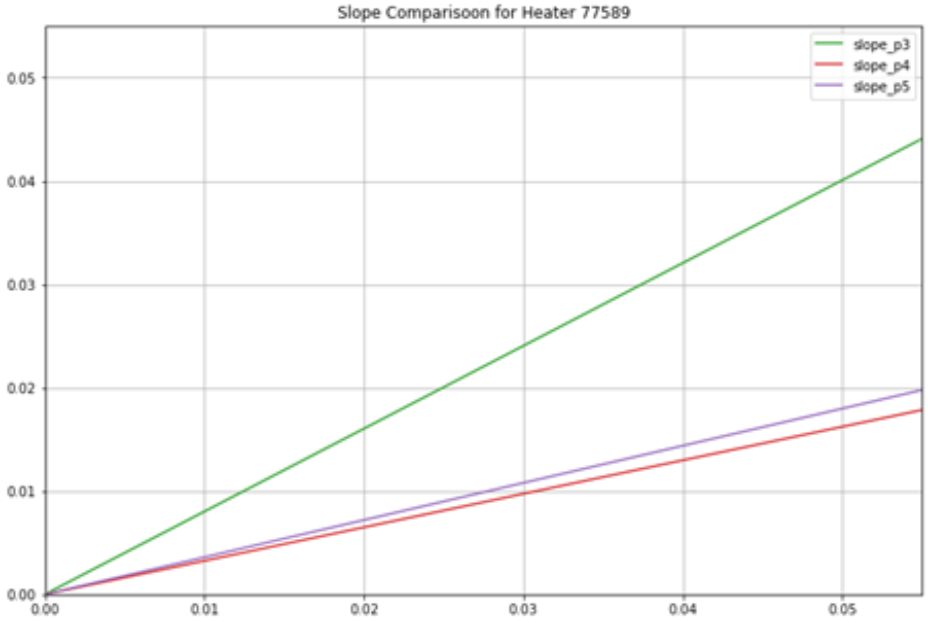

Figure 4.12: Slopes Heater 77589

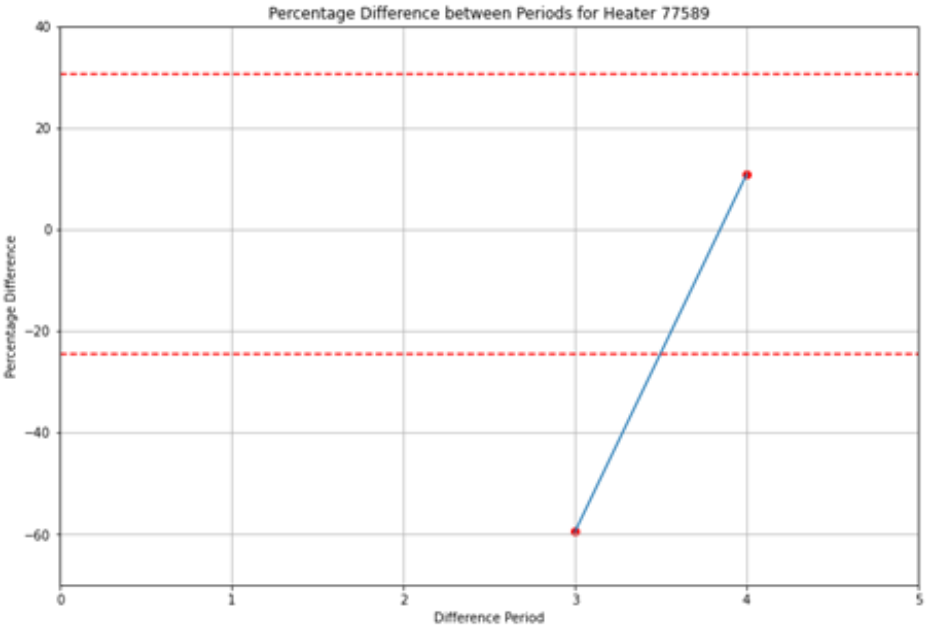

Figure 4.13: Percentage Difference 77589

5 Conclusion and Discussion

With the approach of a support vector regression model, it can be said that a linear model can be used to calculate slopes, from where you can detect significant decreases. With this approach, three complete heating periods must be taken into consideration to make a valid conclusion about a decrease. The previous results show that with additional properties, a filter function can be used to detect heaters where the percentage difference data show a valid decrease and are eligible for a newly added insulation.

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 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.

5.2 Limitations

The model is linear, and you only calculate temperature difference against gas consumption, you better consider multiple parameters and/or non-linear relationships because by only using one parameter in the prediction of the model you see that they are very limited. Since there are more factors that influence gas consumption.

The admissibility of the data is very low because it is not possible to see exactly whether an increase or decrease is due to the gas consumption or is simply due to incorrect data points, changes of the heater itself or other factors.

5.3 Discussion

With the SVR approach, one can therefore very well recognize when there is a decrease in the slope. However, one cannot say with certainty whether this is due to the insulation of the house. The basic question is whether this approach is correct at all in order to determine new insulation on a house on the basis of the data. Although the results can be interpreted to be able to detect this change at all, based on the data available to Intergas or whether other data must be added to make a conclusion about the newly added isolation.

In addition, each household is so individual that a rough filtering based on the density distribution does not take all possible heaters into account. Because an added new insulation of the house can mean a decrease of

-70% for one heater and only -10% for another. However, if the filter is set so low that even a small percentage change in the slope is considered as a possible possibility, the result will not be more accurate and almost all heaters will be affected by the filter. And these, as the density distribution shows, could also just be a normal fluctuation of gas consumption.

5.4 Future Research

For further research, one could cluster the calculated values of the percentage difference of the individual heaters more precisely. With a cluster algorithm for outlier detection, one could say with an exact certainty when an individual heater has a period that shows a significant decrease and thus solve the problem with the density graph based approach and assumption on outlier.

Since Intergas also wants to reduce the amount of data due to performance and early detection which are necessary for the slope calculation, a new approach could be tried to calculate the slopes with as few data points as possible so that they are still meaningful for the respective heating period.

Finally, while investigating the variation in gas consumption, the results showed that there is not only a significant decrease, but also a significant increase in gas consumption. Now the question arises which can be checked in a further investigation where these values come from and whether they are also due to various factors, for example a change in the heater itself or a leak in the supply or system.

Reference

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* .

Geron, A. (2018), *Machine Learning mit Scikit-Learn TensorFlow*, dpunkt.verlag GmbH.

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* .

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer, New York.

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[1]

B | Annotated scripts of analyses and method settings

```
pdf_27729_p1 = df_main[(df_main.heater_id == 27729) & (df_main.period == 1)]
pdf_27729_p2 = df_main[(df_main.heater_id == 27729) & (df_main.period == 2)]
pdf_27729_p3 = df_main[(df_main.heater_id == 27729) & (df_main.period == 3)]
pdf_27729_p4 = df_main[(df_main.heater_id == 27729) & (df_main.period == 4)]
pdf_27729_p5 = df_main[(df_main.heater_id == 27729) & (df_main.period == 5)]
```

Figure B.1: createDataSetsPerHeater-svm2.ipynb[9]

```
# split data for period 1
train_set, test_set = train_test_split(pdf_27729_p1, test_size=0.2,
    random_state=1)
# split data for period 2
train_set2, test_set2 = train_test_split(pdf_27729_p2, test_size=0.2,
    random_state=1)
# split data for period 3
train_set3, test_set3 = train_test_split(pdf_27729_p3, test_size=0.2,
    random_state=1)
# split data for period 4
train_set4, test_set4 = train_test_split(pdf_27729_p4, test_size=0.2,
    random_state=1)
# split data for period 5
train_set5, test_set5 = train_test_split(pdf_27729_p5, test_size=0.2,
    random_state=1)
```

Figure B.2: splitTrainAndTestSetsFromDataset-svm2.ipynb[10]

```
# filter data for period 1
heater_train_data = train_set.copy()
train_gas_use_data = heater_train_data['sum_gas']
train_twoPredictors = heater_train_data[['avg_t_diff']]
heater_test_data = test_set.copy()
test_gas_use_data = heater_test_data[['sum_gas', 'avg_t_diff']] # add
    avg_t_diff here to display on plot
test_twoPredictors = heater_test_data[['avg_t_diff']]
# filter data for period 2
heater_train_data2 = train_set2.copy()
train_gas_use_data2 = heater_train_data2['sum_gas']
train_twoPredictors2 = heater_train_data2[['avg_t_diff']]
heater_test_data2 = test_set2.copy()
test_gas_use_data2 = heater_test_data2[['sum_gas', 'avg_t_diff']]
test_twoPredictors2 = heater_test_data2[['avg_t_diff']]
# filter data for period 3
heater_train_data3 = train_set3.copy()
train_gas_use_data3 = heater_train_data3['sum_gas']
train_twoPredictors3 = heater_train_data3[['avg_t_diff']]
heater_test_data3 = test_set3.copy()
test_gas_use_data3 = heater_test_data3[['sum_gas', 'avg_t_diff']]
test_twoPredictors3 = heater_test_data3[['avg_t_diff']]
# filter data for period 4
heater_train_data4 = train_set4.copy()
train_gas_use_data4 = heater_train_data4['sum_gas']
train_twoPredictors4 = heater_train_data4[['avg_t_diff']]
heater_test_data4 = test_set4.copy()
test_gas_use_data4 = heater_test_data4[['sum_gas', 'avg_t_diff']]
test_twoPredictors4 = heater_test_data4[['avg_t_diff']]
# filter data for period 5
heater_train_data5 = train_set5.copy()
train_gas_use_data5 = heater_train_data5['sum_gas']
train_twoPredictors5 = heater_train_data5[['avg_t_diff']]
heater_test_data5 = test_set5.copy()
test_gas_use_data5= heater_test_data5[['sum_gas', 'avg_t_diff']]
test_twoPredictors5 = heater_test_data5[['avg_t_diff']]
```

Figure B.3: filterValuesFromPreparedTestandTrainSet-svm2.ipynb[11]

```
epsilonValue = 0.5
#train the model period 1
svm_reg = LinearSVR(epsilon=epsilonValue)
svm_reg.fit(train_twoPredictors, train_gas_use_data)
#train the model period 2
svm_reg_p2 = LinearSVR(epsilon=epsilonValue)
svm_reg_p2.fit(train_twoPredictors2, train_gas_use_data2)
#train the model period 3
svm_reg_p3 = LinearSVR(epsilon=epsilonValue)
svm_reg_p3.fit(train_twoPredictors3, train_gas_use_data3)
#train the model period 4
svm_reg_p4 = LinearSVR(epsilon=epsilonValue)
svm_reg_p4.fit(train_twoPredictors4, train_gas_use_data4)
#train the model period 5
svm_reg_p5 = LinearSVR(epsilon=epsilonValue)
svm_reg_p5.fit(train_twoPredictors5, train_gas_use_data5)
```

Figure B.4: trainModelForEachPeriodWithEpsilonValue-svm2.ipynb[12]

```
#validation period 1
gas_use_prediction = svm_reg.predict(test_twoPredictors)
lin_mse = mean_squared_error(test_gas_use_data['sum_gas'], gas_use_prediction)
lin_remse = np.sqrt(lin_mse)
lin_remse
```

Figure B.5: rmseCalculationForPeriod1-svm2.ipynb[13]

```
heaters = df['heater_id'].unique().tolist()
slope = []
for heater_id in heaters:
    df_h = df[df['heater_id'] == heater_id]
    for period in range(1,6):
        df_p = df_h[df_h['period'] == period]
        if(df_p.shape[0]>0):
            x = df_p[["avg_t_diff"]]
            y = df_p[["sum_gas"]]

            svm_reg = LinearSVR(epsilon=0.5)

            model = svm_reg.fit(x, y)
            #print("this is period "+str(period))
            #print(f"slope: {model.coef_}")
            if(period == 1 ):
                if(abs(model.coef_[0]) >= 0.1 ):
                    sp1 = model.coef_[0]
                else:
                    sp1 = None
            elif(period == 2):
                if(abs(model.coef_[0]) >= 0.1 ):
                    sp2 = model.coef_[0]
                else:
                    sp2 = None
            elif(period == 3):
                if(abs(model.coef_[0]) >= 0.1 ):
                    sp3 = model.coef_[0]
                else:
                    sp3 = None
            elif(period == 4):
                if(abs(model.coef_[0]) >= 0.1 ):
                    sp4 = model.coef_[0]
                else:
                    sp4 = None
            elif(period == 5):
                if(abs(model.coef_[0]) > 0.1 ):
                    sp5 = model.coef_[0]
                else:
                    sp5 = None
        else:
            if(period == 1):
                sp1 = None
            elif(period == 2):
                sp2 = None
            elif(period == 3):
                sp3 = None
            elif(period == 4):
                sp4 = None
            elif(period == 5):
                sp5 = None
    L = [heater_id, sp1, sp2, sp3, sp4, sp5]
    slope.append(L)
```

Figure B.6: calculatingSlopesForEachHeaterPeriod-svm_final.ipynb[6]

```
cols = ['heaterid', 'slope_p1', 'slope_p2', 'slope_p3', 'slope_p4','slope_p5']
slope_df = pd.DataFrame(slope, columns=cols)
slope_df = slope_df.apply(pd.to_numeric)
```

Figure B.7: calculatingSlopesForEachHeaterPeriod-svm_final.ipynb[7]

```
def differences (row) :
    if(row['slope_p1'] == None or row['slope_p2'] == None):
        row['diff_1'] = None
    else:
        row['diff_1'] = row['slope_p2'] - row['slope_p1']
        if(row['slope_p1'] == 0):
            row['pdiff_1'] = None
        else:
            row['pdiff_1'] = (row['diff_1']/row['slope_p1'])*100
    if(row['slope_p2'] == None or row['slope_p3'] == None):
        row['diff_2'] = None
    else:
        row['diff_2'] = row['slope_p3'] - row['slope_p2']
        if(row['slope_p2'] == 0):
            row['pdiff_2'] = None
        else:
            row['pdiff_2'] = (row['diff_2']/row['slope_p2'])*100
    if(row['slope_p3'] == None or row['slope_p4'] == None):
        row['diff_3'] = None
    else:
        row['diff_3'] = row['slope_p4'] - row['slope_p3']
        if(row['slope_p3'] == 0):
            row['pdiff_3'] = None
        else:
            row['pdiff_3'] = (row['diff_3']/row['slope_p3'])*100
    if(row['slope_p5'] == None or row['slope_p4'] == None):
        row['diff_4'] = None
    else:
        row['diff_4'] = row['slope_p5'] - row['slope_p4']
        if(row['slope_p4'] == 0):
            row['pdiff_4'] = None
        else:
            row['pdiff_4'] = (row['diff_4']/row['slope_p4'])*100

    return row
slope_df = slope_df.apply(lambda row: differences(row), axis=1)
```

Figure B.8: calculatingSlopesForEachHeaterPeriod-svm_final.ipynb[8]

C | Full analysis results

	_c0	heaterid	slope_p1	slope_p2	slope_p3	slope_p4	slope_p5
1738	1738	27729.0	0.485647	0.468805	0.480826	0.191385	0.212005
7713	7713	57721.0	NaN	0.414850	0.426396	0.275337	0.290224
7790	7790	8180.0	0.692461	0.717407	0.414945	0.468698	0.515741
8327	8327	45441.0	NaN	1.150599	0.583266	0.553134	0.634685
8812	8812	77589.0	NaN	NaN	0.801614	0.324421	0.359556

Table C.1: Complete final results preview part 1

diff_1	pdiff_1	diff_2	pdiff_2	diff_3	pdiff_3	diff_4	pdiff_4
-0.016842	-3.467938	0.012021	2.564252	-0.289442	-60.196727	0.020621	10.774495
NaN	NaN	0.011546	2.783179	-0.151059	-35.426834	0.014887	5.406702
0.024946	3.602508	-0.302462	-42.160428	0.053753	12.954196	0.047043	10.036981
NaN	NaN	-0.567332	-49.307566	-0.030132	-5.166152	0.081551	14.743442
NaN	NaN	NaN	NaN	-0.477194	-59.529076	0.035135	10.830167

Table C.2: Complete final results preview part 2

```
1 def filterHeatersWithValidDecrease(heater_df):
2     heater_df_ZN = slopes_df[((slopes_df.slope_p1 > 0) | np.isnan(slopes_df.
          slope_p1)) & ((slopes_df.slope_p2 > 0) | np.isnan(slopes_df.slope_p1)) &
           (slopes_df.slope_p3 > 0) & (slopes_df.slope_p4 > 0) & (slopes_df.
          slope_p5 > 0)]
3     final_heater_result = heater_df_ZN[((heater_df_ZN.pdiff_1 < -24.5) & (
          heater_df_ZN.pdiff_2 < 30.5) & (heater_df_ZN.pdiff_3 < 30.5) & (
          heater_df_ZN.pdiff_4 < 30.5)) |
4                                        ((heater_df_ZN.pdiff_2 < -24.5) & (
                                             heater_df_ZN.pdiff_3 < 30.5) & (
                                             heater_df_ZN.pdiff_4 < 30.5) & ((
                                             heater_df_ZN.pdiff_1 < 30.5) | np.
                                             isnan(slopes_df.slope_p1))) |
5                                        ((heater_df_ZN.pdiff_3 < -24.5) & (
                                             heater_df_ZN.pdiff_4 < 30.5) & ((
                                             heater_df_ZN.pdiff_1 < 30.5) | (np.
                                             isnan(slopes_df.slope_p1))) & ((
                                             heater_df_ZN.pdiff_2 < 30.5) | (np.
                                             isnan(slopes_df.slope_p2))) )]
6     return final_heater_result
```

Figure C.1: filterHeatersWithValidDecrease-possibleHeaterDetection.ipynb[5]

	_c0	heaterid	slope_p1	slope_p2	slope_p3	slope_p4	slope_p5
10211	10211	8941.0	0.433552	0.53911	0.551414	0.520482	0.55422

Table C.3: Complete data for heater 8941 part 1

diff_1	pdiff_1	diff_2	pdiff_2	diff_3	pdiff_3	diff_4	pdiff_4
0.105557	24.347091	0.012305	2.282381	-0.030932	-5.60961	0.033738	6.482098

Table C.4: Complete data for heater 8941 part 2