UTRECHT UNIVERSITY

Department of Information and Computing Sciences

Applied Data Science Master Thesis

Finding human behavioural in heat pump power usage

Using Fourier Transform Analysis

First examiner:
Laurens Stoop

Second examiner:
Ad Feelders

Candidate:
Sahar Pourahmad

In cooperation with:
Inversable BV
Intergas BV

July 9, 2023

Abstract

Due to the increasing environmental challenges brought about by climate change, the Netherlands is transitioning to a carbon-neutral economy. As part of this transition, the country aims to reduce CO_2 emissions and improve energy efficiency, particularly in the housing sector. Hybrid heat pumps have been identified as a potential solution. However, further investigation is needed to assess the performance of hybrid heat pumps and their suitability for different homes.

Currently, the *Inversable Demo Project Hybrid* is being conducted in the Netherlands to provide insights to the Dutch government regarding the investment in hybrid heat pumps and the potential for their wider adoption in the future. This thesis focuses on analysing energy usage across houses and exploring potential societal or human behavioural patterns to improve the prediction and understanding of heat pump power usage.

Fourier transform was utilized in conjunction with linear regression to reveal hidden frequent patterns in heat pump power usage, with the objective of exploring their relationship with indoor and outdoor temperatures and the possibility to find a general pattern for heat pump power usage. Although some indications of recurrent patterns were observed, conclusive and definitive evidence of societal or human behavioural patterns could not be established. Further investigation is needed to validate these findings and explore the effects of behavioural differences on hybrid heat pump performance.

Preface

This master's thesis was undertaken as a part of the Applied Data Science program at Utrecht University. Two teams of master students, collaborated with Inversable BV and Intergas Verwarming BV, which provided the team with the dataset collected for the *Demoproject Hybride*. This project, initiated in November 2021, aims to gain valuable insights into the practical performance, savings, and applicability of hybrid heat pumps.

This thesis is based on the combined work of Sahar Pourahmad and Ruben de Groot. The overall aim was to find human influence within the temporal component of hybrid heat pump power usage. The methods applied by both students are interwoven and built on top of each other. This specific thesis focuses on finding patterns in the time series through using Fourier transform, while the other project uses predictive modelling to establish a relationship and model the overall behaviour of heat pump energy usage. The combination of methods gives an insight into the possibilities of detecting patterns influenced by human behaviour and how this relates to differences in heat pump power usage between different devices.

We express our gratitude to Inversable BV and Intergas Verwarming BV for providing us with access to the dataset, which formed the cornerstone of our investigation. Furthermore, we extend our appreciation to our supervisors and mentors who offered their guidance and expertise throughout this research endeavour. Jordi Beunk, Johanna Lems, Abdulhakim Özcan, Sahar Pourahmad, and Ruben de Groot collectively dedicated their time, skills, and efforts to this study, and we are proud to present the findings of our research in this master's thesis.

Special thanks

Special thanks are extended to my supervisors Laurens Stoop and Erwin Bisschop, as well as my colleague Ruben de Groot, for their support and contribution throughout this journey.

Erwin, as a representative of Inversable, provided access to the dataset, shared the necessary information and domain knowledge with us, and patiently answered all of our questions.

Laurens's valuable insights, guidance, ideas, and dedicated time to this project are highly appreciated and have played a crucial role in its success.

Finally, a special mention goes to Ruben for his significant contributions to the data extraction for the whole group of Utrecht University students working with Inversible. In addition, his substantial efforts in data wrangling, preparation, and predictive modelling in this specific thesis cannot be neglected.

This thesis would have been nowhere without the help of these special individuals.

Contents

1. Introduction

1.1 Climate change

It is undeniable that the Earth and its entire population are confronted with rapid and alarming environmental challenges. In the past century, human activities have produced an artificial increase in the concentration of greenhouse gases in the atmosphere, causing the trap of the sun's energy in the earth's system. According to NASA the average surface temperatures of Earth is estimated to rise between 2°C and 6°C by the end of the 21st century and the rate of global warming has nearly doubled in the last 50 years [1].
The impact of global warming is far greater than just increasing temperatures. It has disrupted the natural water cycle, resulting in more intense rainfall, flooding, and drought in various regions. Global warming also affects rainfall patterns and causes rising sea levels, coastal erosion, and shifts in infectious disease ranges. The impact of rising greenhouse gas emissions on climate change is already evident and requires urgent action [1][2].
By 2019, the concentrations of atmospheric carbon dioxide (CO_2) had reached levels higher than those observed in at least the past 2 million years [2]. As a result, international measures are being implemented to tackle climate change and its negative impacts. In 2015, world leaders of 194 parties, including the European Union, joined the United Nations Paris Agreement, committing to work together towards a net-zero emissions world [3].

1.2 Energy transition in the Netherlands

Among the countries of the European Union, the Netherlands has made notable progress on its transition to a carbon-neutral economy. The country is aiming for a swift transition to a low-carbon economy and has integrated greenhouse gas reduction targets into its energy and climate policy. In 2019, a Climate Agreement package was developed by the business community and civil society organizations [4][5] aiming to reduce CO_2 emissions by 49% in 2030 compared to 1990 [6].
One of the key measures outlined in the Climate Agreement focuses on enhancing the energy efficiency of homes, and transition away from natural gas heating for new buildings, while also urging improvements in existing buildings to enable fossil-free heating methods [7].
A recent report from the Dutch Heating Industry (NVI), confirms that the CO_2 reduction target in the built environment by 2030 can be accomplished by installing 1.7 million hybrid heating systems. This report emphasizes the potential of hybrid heat pumps, as a suitable and sustainable solution for millions of homes [8]. In May 2023, the government stated that there would be stricter requirements regarding the efficiency of heating installations, and heat pumps will become the minimum standard starting from 2026 [9].

1.3 About Heat pumps

A heat pump can either serve as a replacement for the central heating boiler, or work in conjunction with it. It can heat the house by taking the heat from the outside air, or other sources such as soil or solar panels, providing an energy efficient heating solution [10].
Heat pumps come in various types, with the primary differentiation being between fully electric heat pumps and hybrid heat pumps. A fully electric heat pump provides heating for the house, and for the hot water in the kitchen and bathroom. However, this is only a suitable choice for the houses that are reasonably to well insulated. In cases where the insulation of the house is not optimal, a hybrid heat pump can be used.
A hybrid heat pump consists of an electric heat pump and a central heating boiler. The heat pump plays a significant role in providing heat for the house, while the central heating boiler is activated during extremely cold weather and for heating tap water [11]. Figure 1.1 visually illustrates the components of a heat pump system, which consist of an outdoor unit with an electric heat pump, an indoor unit featuring another electric heat pump, a boiler, a hot water provider, and a heat pump heating unit. Additionally, Figure 1.2 provides a real-life photograph of a heat pump, offering a visual representation of the system's physical appearance.

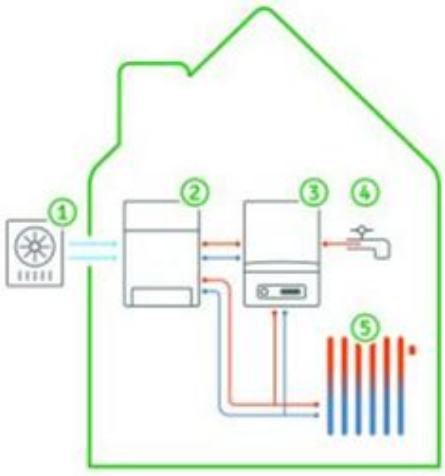


Figure 1.1 - The components of a heat pump system includes an outdoor electric heat pump (1), an indoor electric heat pump (2), a boiler (3), a hot water provider (4), and a heat pump heating unit (5) [10].

(a) Hybrid heat pump (left) with central heating boiler (right)

(b) The outdoor unit

Figure 1.2 – a real-life photograph of a heat pump [11]

A heat pump works as follows:

1. Heat is obtained from the air outdoors and directed to the heat exchange surface of the outer component of the heat pump.
2. The heat makes the liquid refrigerant inside the heat pump evaporate and transform into a gas.
3. This gas is then transported through a compressor, which boosts its pressure, resulting in an increase in temperature.
4. The heated gas is directed over the internal heat exchange surface. This heat can be either blown throughout the interior of the house or transferred to a central heating or hot water system.
5. As the heat is transferred into the house, the gas cools down, causing it to revert to a liquid state.
6. This cycle of reverse refrigeration repeats until the desired temperature is reached in the house, as set on the thermostat.

One advantage of the hybrid heat pumps is their versatility, that allows them to be installed in various housing types with minimal adjustments to the living situation. As most houses already have a boiler and so the needed pipes and connectors are already in place, the installation is generally not a big operation. Also, it has been confirmed that by implementing a hybrid heat pump, a reduction of around 20 percent in CO_2 emissions from heating and hot water can be achieved, accompanied by an approximate decrease of 80 percent in natural gas consumption [10].

1.4 Previous research

There have been earlier projects about monitoring the performance of heat pumps in the Netherlands.

In 2019 the project *Installatiemonitor* started. The goal of this project was to get information about the real-life performance of both heat pumps and hybrid heat pumps. The project was a partnership between Enpuls, Gasterra, Gasunie, Liander, N-Tra, RVO, Stedin and Techniek Nederland and was carried out by consultancy firm BDH. During this project, 800 heat pumps were monitored until the 30th of June 2021, from which 450 were eventually analysed. They concluded that the release temperature of the heating system and surface of energy loss positively correlated with the energy usage of the heat pump and that a hybrid heat pump reduces CO_2 significantly and that a hybrid heat pump is a financially attractive option [12].

1.5 Demo Project Hybride

Currently, in the Netherlands, a subsidy is available for the purchase of hybrid heat pumps and the Dutch government is actively gathering further insights on the real-life performance of these hybrid heat pumps through the *Inversable Demo Project Hybride*. The Inversable Demo Project Hybride is a collaborative effort involving several organizations and institutions, including the Dutch Heating Industry (NVI), Technology Netherlands, Ministry of the Interior and Kingdom Relations (BZK), Ministry of Economic Affairs and Climate Policy (EZK) and the Netherlands Enterprise Agency (RVO). In addition, Utrecht University is also participating in this project. The current manufacturers participating in this project are Atag, Ferroli, Intergas, Nefit Bosch, Remeha and Vaillant. The project involves monitoring hybrid installations in around 200 homes for at least one heating season, and analysing the applicability, performance, savings, and comfort of hybrid heat pumps [10].

1.6 Research question

Currently, Inversable has identified all 200 participants and continues to expand the data collection as new heat pumps are installed and get connected to a database. Although lots of data has been collected by 30 different sensors and analysed until now, there remains unresolved inquiries that need further investigation, including the potential human behavioural patterns in heat pump performance. It is likely that energy usage varies among different houses, based on the type of the house, energy label, and the number of occupants. However, the focus of this thesis is to investigate whether there is a general energy usage pattern that can explain the behaviour of all houses. Additionally, this research aims to explore whether there are any societal or human behavioural patterns that can improve the ability to predict and understand heat pump power usage more effectively.
Based on the research conducted within the Inversable Demo Project Hybrid, a report will be presented to the Dutch government that is about the evaluation of the performance of hybrid heat pumps, in order to help them decide whether this worths the investment. If the demo proves successful, it will lead to the installation of more hybrid heat pumps in the future.

1.7 Thesis outline

The remaining sections of this thesis are organized as follows:

Section 2 will provide details on the data, including its specification, extraction, preparation, and a brief data analysis. Section 3 will dive into the methodology used in the research. The results will be presented in Section 4. Finally, Section 5 will present the conclusions drawn from the study and provide suggestions for potential areas of future research.

2. Data

This chapter includes a description of the data followed by data extraction, selection, explation and data cleaning.

2.1 Data description

As highlighted in section 1.6, this project involves 200 participants. However, the time-series data comprises real-time measurements recorded in 169 houses, as not all of them have their heat pump installed yet. On average, each house has approximately 7 months of measured data, with the duration ranging from 20 days for the shortest period to 15 months for the longest. During this period, various of sensors were deployed in the houses, including the heating system sensor, heat pumps, boilers, smart meter, and the indoor climate sensor. In addition, the sensors within the homes and local weather data were also collected for further analysis.

A summary of these sensors is provided in Table A-1, Appendix A. Here is a brief description of each sensor:

- Heating System Sensor: Monitors the heating system within the house, measuring parameters such as flow rate, supply and return temperatures. Energy and power of the thermal system are derived from these measurements. The heating system sensor provides detailed insights into the heating system's performance, collecting data every 5 seconds.
- Heat Pump Sensor: Positioned at the heat pump unit, it measures the energy and power supplied to the heat pump. However, it does not directly measure the amount of energy converted by the heat pump and transferred to the heating system.
- Boiler Sensor: Monitors energy and electricity usage of the boiler. Note that a portion of this energy may be used for purposes other than heating systems, such as water heating for showers.
- Smart meter: Reads smart meter data for the house, transmitting power readings every 60 seconds and energy and gas readings every 10 seconds. This enables real-time monitoring and analysis of energy consumption patterns. The power consumed and delivered values represent the combined total for all three phases of the network. Similarly, the energy consumed and delivered accounts for both high and low tariffs. Energy can be delivered by a house when it generates its own energy, for instance, through the utilization of solar panels.
- Indoor Climate Sensor: Typically placed in the living room, it measures indoor temperature and humidity, providing valuable data for understanding indoor climate conditions.

- Local Weather Data: Collected from The Royal Netherlands Meteorological Institute (KNMI) and linked with each house by Inversable, providing additional information on weather conditions.

Although metadata of each house was also provided, this was not used in this research and therefore not described.

2.2 Data extraction

To extract the necessary data, queries were formulated and executed against the Influx database of Inversable. The query results were then retrieved and converted into Python Pandas data frames. The data used in this thesis was obtained from three sources: the heat pump, the indoor climate sensor, and the KNMI (The Royal Netherlands Meteorological Institute). Specifically, the heat pump data provided power measurements in watts, the climate sensor data included inside temperature readings in degrees Celsius, and the KNMI data offered outside temperature readings in degrees Celsius.

2.3 Data selection

A list of active devices given by Inversable was used to see what devices were usable. Devices were not activated at the same time, meaning that they were not well comparable. For instance, if one device was active during the winter of 2022 and the other one was not, they would have a different average heat pump power usage over the whole period. This was important because it could influence the normalized heat pumper if a Z-score normalization was applied. Figure 2.1 illustrates the monthly increase in the number of active devices, highlighting the varying levels of device activation over time.

The data analysis focused on the autumn to spring period, which corresponds to the period of highest heating usage. To ensure consistency in the analysis, the data selection ranged from October 6th 2022, to April 1st 2023. This duration encompassed 59 active devices.

Per device, the data was aggregated to daily values within the query procedure. In addition, only the period the device was active was extracted, further reducing the query size.

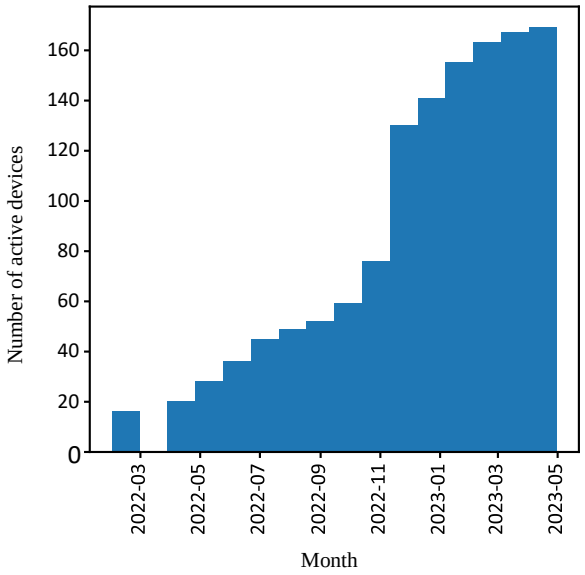


Figure 2.1 – Number of active devices within the Demo Project Hybrid at each month

The variables that are used in this thesis are aggregated as follows: Heat pump power is summed per day and for the inside and outside temperature the daily mean was used. Figure 2.2 shows an illustration of one of the houses plotted against dates.

For the local weather data, besides aggregation, the values were multiplied by 10 because the KNMI recorded the data in a scaled format, where the value '0.1' corresponded to a magnitude of 1, and the value '2' represented a magnitude of 20 °C.

All values were then normalized using Z-score normalization to enable fair comparisons between devices. This involved subtracting the mean and dividing by the standard deviation of each value.

2.4 Data exploration

In this section, a brief data exploration is conducted on the mean values of each month, for heat pump power, temperature inside, and temperature outside of the house. Figure 2.3 illustrates the mean values for each variable from October 2022 to March 2023. The analysis reveals that the heat pump power usage has an average around 500 W throughout the period, with the lowest average in October and the highest averages in December to February (figure 2.3-a). Conversely, the temperature outside of the house exhibits the opposite pattern, with an average of 7 degrees Celsius during this period (figure 2.3-c). Additionally, the temperature inside the house demonstrates a relatively constant mean value of approximately 20 degrees Celsius in all months (figure 2.3-b).

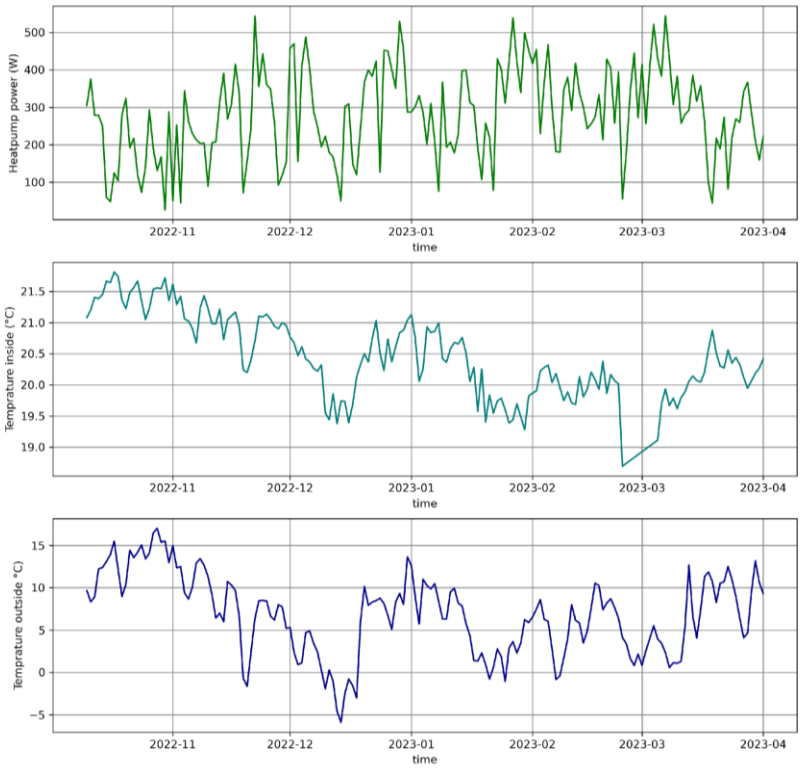
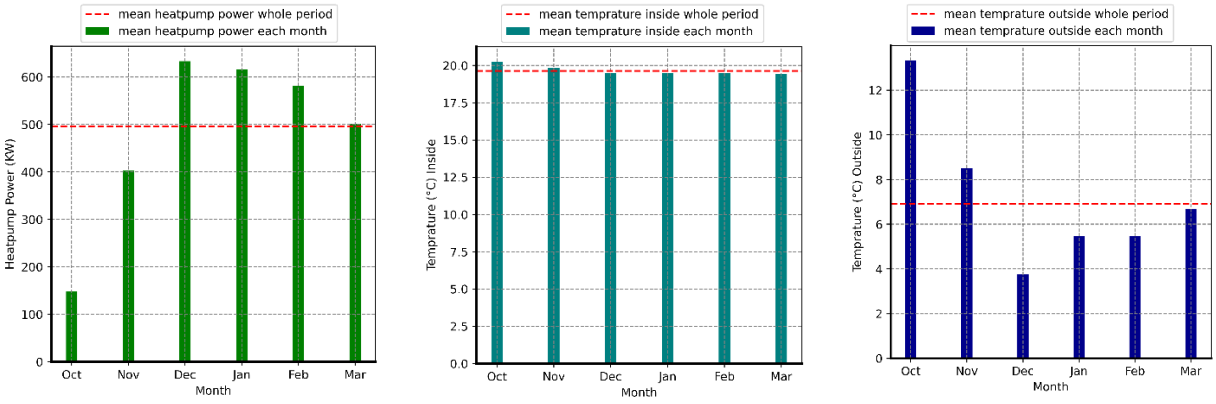


Figure 2.2 – A timeseries plot of daily power usage, the temperature inside and outside of house id 'z17bDdY4'



a) heat pump power b) temperature inside c) temperature outside

Figure 2.3– mean values of variables for each month and during the whole period of 6 month

The opposite relationship between the heat pump power and temperature outside of the house was further investigated by plotting the mean values of each variable at each month in Figure 2.4. The plot reveals a relatively linear behaviour with a negative slope, confirming the inverse relationship

between heat pump power and temperature. This can mean that colder weather conditions require more energy consumption for heating purposes, leading to higher power usage. This is an interesting finding and should be considered during the rest of the analysis and interpretation of results.

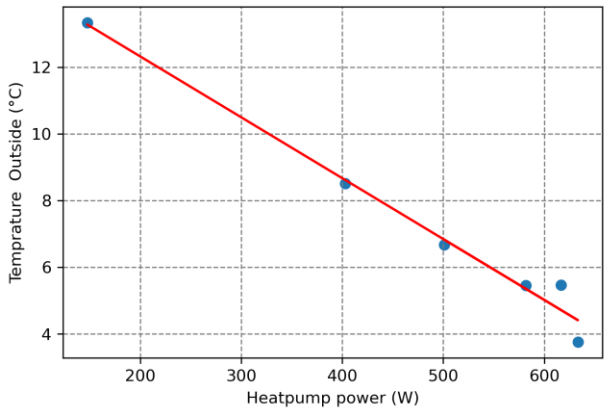


Figure 2.4 - Mean values of heat pump power and temperature outside of the house at each month

2.5 Data preparation

Real-world data can often be incomplete, noisy, and inconsistent. The data of this project was not an exception, therefore, to address these issues, data cleaning techniques were applied during the pre-processing stage prior to model deployment. The goal of data cleaning was to fill in missing values, reduce noise by identifying outliers, and correct inconsistencies in the data.

2.5.1 Outliers

An outlier is a data object that significantly deviates from the majority of the dataset. In contrast to noisy data, which represents random errors or variance, outliers are suspected of being generated by different mechanisms compared to the rest of the data. When an outlier is suspected to be a result of data collection or recording error, one possible approach is to remove the observation [13] [14].

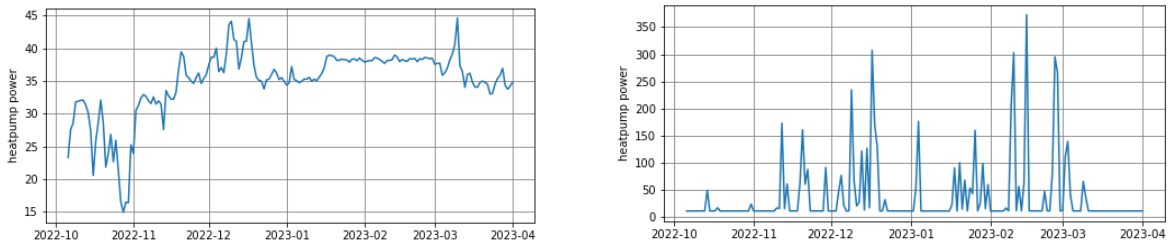
Detecting outliers in a dataset that is influenced by human behaviour presents special challenges. Even though the outliers do not follow a recurring pattern, they may still be attributed to human behaviour, such as activities during holidays or special events. However, it is important to note that occasional behaviours that occur only once or twice a year may not be easily recognizable in a dataset that covers only around 8 months of heat pump usage.

Outliers in the original data were detected by utilizing a threshold of 2.5 times the standard deviation. The selection of the appropriate coefficient of standard deviations was determined through an iterative process involving trial and error. Initially, using a standard deviation of 3 resulted in very few values being identified as outliers, even though there was evidences of some being present. Conversely, using a standard deviation of 2 classified too many values as outliers, including those that likely represented the human behaviour patterns or spikes caused by temperature changes. By selecting a threshold of 2.5 times the standard deviation, we aimed to strike a balance and identify outliers that were statistically significant while considering the potential influence of human behaviour and temperature fluctuations.
In the end, only a few data points per device were identified as outliers, almost all of which used more heat pump power. This was expected, because the beginning of autumn has very low heat pump power usage values, close to zero, meaning that heat pump power values on the lower end are almost never marked as an outlier. After finding the positions of the outliers, their values were transformed into nan values.

2.5.2 Removing strange devices

In some cases, whole times series measurements related to a house was dropped because of having a strange values altogether. An example of one of these devices is visible in figure 2.5. The average heat pump power usage of the house in figure 2.5-a is too low compared to most other houses and suggests that another form of heating is used. Also, the strange behaviour of the house in figure 2.5-b can indicate a faulty measurement or another form of heating as well. As Inversable added a list of devices that exhibited similar odd behaviour this list combined with our own research, led to the deletion of 8 devices, leaving 45 remaining devices.

Furthermore, some devices showed a different behaviour during freezing temperatures. As a second cleaning stage for the creation of the final model in section 4.2, 15 of these devices were removed. Therefore, the data set used in that part only contains devices that use more heat pump power than average during freezing temperatures.



(a) (b)

Figure 2.5 – heat pump power is house id (a) ol6oCqA4 and (b) xl6jHNhD

2.5.3 Imputation of missing data

Datasets often contain missing values, which can present challenges as many statistical learning methods cannot directly handle missing values. An appropriate strategy should be decided to address the missingness.

One option is to remove the rows that contain missing observations and only analyse the complete rows. This technique, however, may be wasteful and unrealistic, in addition to compromising data consistency. Another common strategy is to impute the missing values and create a completed matrix that can be used in statistical learning methods [13][14].

There are several ways for imputing the missing values, such as mean imputation, regression imputation, and stochastic regression imputation. Regression Imputation was chosen in this project to impute both missing data and the removed outliers as it has the advantage of using the information from other variables to generate smarter imputations for the missing data.

The process of imputation with regression involves constructing a model based on the observed data, and then using this model to predict values for the incomplete cases, which can serve as replacements for the missing data. However, this method has limitations as well: the variability of the imputed data is systematically underestimated, and it tends to bias the correlations between variables upwards, meaning that the estimated relationships between variables tend to be stronger than they are [15].

In this project a linear relation between temperature and heat pump power was found. However, the sub-zero temperatures showed large deviations from the estimated relation. After removing the odd devices, and the outliers and imputing the data, a few devices had still missing values at the beginning of the extracted period. Therefore, we made the decision to shorten the period by moving the start date a few days forward, changing it from October 6 th to10 th 2022.

3. Methods

As mentioned earlier, this research aims to investigate the presence of societal and human behavioural patterns in heat pump power in houses participating in the demo hybrid project. To address this matter, two main statistical methods were used, namely Fourier transform and multiple linear regression. While the focus of this thesis is on the Fourier transform, it is important to note that the application of Fourier transform analysis heavily relied on the linear regression model. In this section both of these method are going to be discussed.

3.1 Fourier transform

Fourier transform, originally coming from the field of digital signal processing [16] is a powerful mathematical technique which is used to analyse the signals in the frequency domain, by decomposing them into the sum of simpler sinusoidal signals [17] (figure 3.1).
Fourier transform applications are beyond its original domain and it is commonly used in the analysis of time series data due to their common characteristics with signals [16]. Time series data consist of measurements recorded at discrete time points [14] making them temporal dependent like signals. This temporal dependency indicates that the data carries information throughout time, and it can contain periodic patterns [18].

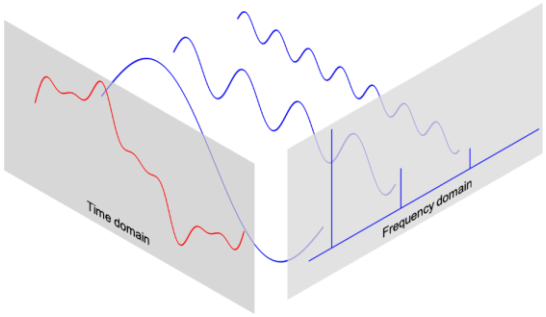


Figure 3.1 – Fourier transform analyses the signals in the frequency domain, by decomposing them into the sum of simpler sinusoidal signals [21]

By applying Fourier analysis on signals, the dominant frequencies within the signals are extracted, along with their relative amplitudes and phased. Similarly, Fourier analysis on time series data, can unveil the hidden underlying periodic patterns present in a time series data, providing insights into their frequency and significance [19][20].

The Fourier transform is a mathematically complex concept, and diving into its details is out of scope of this thesis. Instead, a simplified explanation of the Fourier transform will be provided, focusing on its practical concepts in an applied manner.

However, it is important to first introduce the basics of signals to establish a foundation for understanding how to model and study them.

3.1.1 Basics of a signal wave

A wave can be defined as a form of periodic motion that repeats at regular intervals [22]. Here we are interested in a specific type of periodic motion known as simple harmonic motion (SHM), that is can be written as a sine or cosine function of time. Formula 1 defines the behaviour of such a motion in time, where A is the amplitude, t is the time in seconds (s), ϕ is the phase constant in radians, and f is the frequency in Hertz (Hz).

Formula 1 – Simple Harmonic Motion based on time $$x(t) = A\cos(2\pi ft + \phi)$$

The amplitude (figure 3.2), is the difference between the peak and the baseline value, and indicate the significance of a wave's variation. A larger amplitude suggests a more influential wave, while a smaller amplitude indicates a weaker wave.

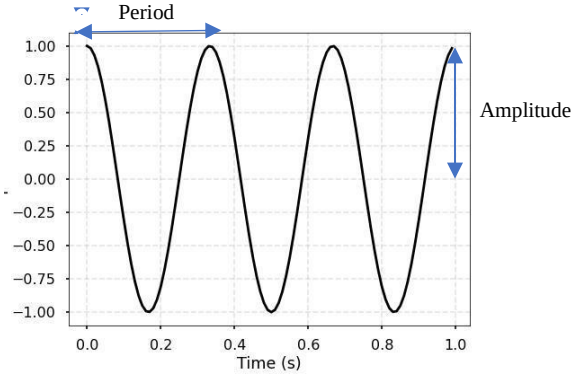


Figure 3.2 – The amplitude and the period of a wave

The phase constant of a signal represents the position of the waveform within a cycle, and it is typically measured in radians. While a complete cycle of a sinusoidal wave corresponds to an angle of 2π radians (360 degrees), the phase of a signal represents the angular distance from the reference point.

While period is the time that it takes for one full oscillation to finish in seconds (figure 3.2), frequency refers to the number of cycles that occur within one second and is measured in Hertz (Hz) [21][22]. The relationship between period (T) and frequency (f) is inversely proportional can be expressed using the formula 2:

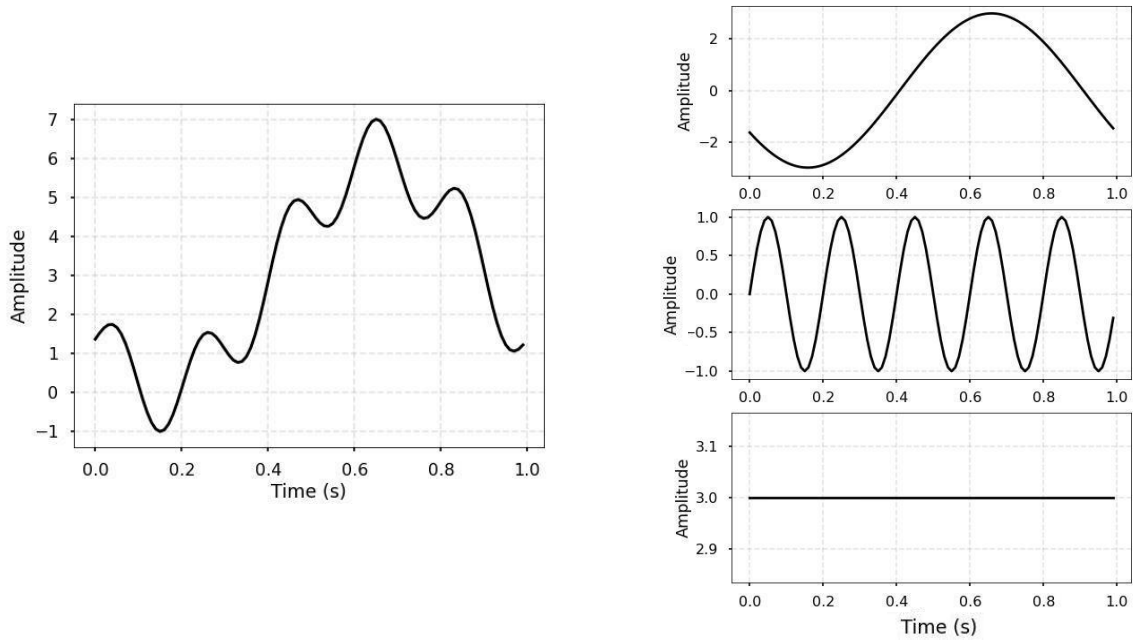
Formula 2 – The relationship of frequency and period $$T\ (s) = \frac{1}{f\ (Hz)}$$

With this basic introduction of waves provided, let's explore the mechanism of the Fourier transform with a simple and straightforward example.

3.1.2 An example of Fourier transform method

Consider a complex signal (Figure 3.3-a) composed of three simple sinusoidal waveforms (Figure 2.2-b):

1. The first sinusoidal waveform oscillates at a frequency of 1 Hz with an amplitude of 3 units and a phase shift of 10 radians.
2. The second sinusoidal waveform has a higher frequency of 5 Hz, an amplitude of 1 unit, and a phase of 0 radians.
3. The third component is a constant waveform that remains steady at a value of 3 units.



a) b)

Figure 3.3 – A complex signal (a) that is composed of 3 sine waves (b)

When the individual waveforms that contribute to the complex signal are unknown, the Fourier transform can be used to determine the specific components that formed the wave, by breaking it down to its parts and identify the underlying components responsible for its formation. Figure 3.4 shows all the frequency components present in the original signal, that were identified using Fourier transform method.

The Fourier transform analysis of the complex signal shows that it is comprised of the summation of three distinct waves, with the frequency of 0, 1 and 5 Hz and the amplitude of 0.5, 1.5 and 3 units, respectively.
It is important to mention that the frequency of zero indicates a constant component in the signal and does not contribute to any oscillatory patterns. In other words, it means that the signal has a steady component over time and can represent the average value of the variable. Also, the negative frequencies that appear in the Fourier transform output, are because of the mathematical features of the Fourier transform. However, in many applications, including this project, negative frequencies do not provide meaningful information and are therefore disregarded.

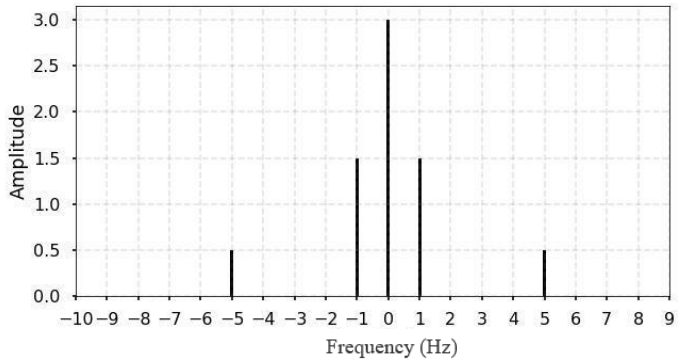


Figure 3.4 - The x-axis represents the frequency in Hertz (Hz), while the y-axis represents the amplitude, which indicates the strength or magnitude of each frequency component.

In order to apply the Fourier transform and comprehend the results, it is essential to look into the mathematical foundations of this technique. Next section will provide a short overview of the Fourier transform.

3.1.3 Basics of Fourier transform

The one-dimensional Discrete-time Fourier transform (DFT) technique has been chosen in this project due to the data's one-dimensional nature (heat pump power) and regular discrete time sampling [19]. Specifically, the Fast Fourier Transform (FFT) was used in this project, as a fast algorithm for DFT. FFT takes advantage of symmetries in the DFT to reduce computational complexity and was preferred due to its simplicity, speed, and ease of implementation [23].

The DFT provides information about the frequency components that comprise the original signal by decomposing it into a sum of sine waves, and is defined in formula 3:

Formula 3 - One-dimensional Discrete-time Fourier transform

$$X_k = \sum_{k=0}^{N-1} x_n . e^{-i2\pi kn/N}$$

Where N represents the total number of samples, n represents the current sample index, and k represents the current frequency index ranging from 0 to N-1. In this formula, e is Euler's number, which is approximately equal to 2.7. Also X_k is a complex number that encodes both the amplitude and phase of a complex sinusoidal component $e^{-i2\pi kn/N}$, present in original signal x_n [21][24].

By applying Euler formula (formula 4), a connection between the exponential function and trigonometric functions is established, resulting in formula 5:

Formula 4 – Euler equation

$$e^{ix} = cos\ cos\ x + i\ sin\ sin\ bx$$

Formula 5 – DFT formula derived by Euler equation

$$X_k = \sum_{k=0}^{N-1} x_n \left[cos\ \left(\frac{2\pi kn}{N}\right) - i.sin\ \left(\frac{2\pi kn}{N}\right)\right]$$

The Fourier coefficients of a time series data can be calculated as complex numbers using the FFT library in SciPy. The frequencies present in the data can be retrieved using the fft.fftfreq function from the same library. Furthermore, the amplitude (A) and phase (ϕ) of each identified wave can be determined through formula 6:

Formula 6 – Calculating amplitude and phase of a signal from Fourier coefficients

$$A = \frac{\sqrt{Re(X_k)^2 + Im(X_k)^2}}{N}$$

$$\phi = \text{atan2}(\text{Im}(X_k), \text{Re}(X_k))$$

Where *Im*(*Xk*) and *Re*(*Xk*) are the imagery and real part of the complex number, *atan*2 is the two-argument form of the *arctan* function [19].
The frequency and amplitude components present in a time series data provide valuable insights into recurring patterns within the data. We can determine the periods at which these patterns occur, measured in the unit of time, using the formula 2. Additionally, by choosing the largest amplitudes, we can identify the most significant patterns within the time series. These calculations enable us to gain a deeper understanding of the most impactful recurring patterns in the data.

On the other hand, as discussed in section 3.1.1, the behaviour of a signal can be characterized, and the corresponding waves can be generated using formula 1. By obtaining the frequency, amplitude, and phase information of a signal, we can regenerate the constituent waves by focusing on the most significant components that comprise it. By employing the regenerated waves from multiple houses in a linear regression model, we can identify coefficients associated with each house. This approach allows us to introduce a general pattern for the time series data, along with the varying factors specific to each house. By successfully achieving this, we can gain valuable

insights into the overall behaviour of the time series data and uncover unique behavioural patterns specific to each house.

3.2 Linear regression

Linear regression is a supervised learning method that is widely used for predicting the outcomes or analysing the association between variables. The first assumption of this method is that there exists a linear relationship (formula 7) between the predictor variables (x_i) and the response variable (Y).

formula 7 – Multiple linear regression

$$Y = \beta 0 + \beta 1 \quad x1 + \beta 2x2 + \cdot \cdot \cdot + \beta pxp + \epsilon$$

Where β_i are unknown constants, representing the model coefficient, and ϵ is a mean-zero random error term. We typically assume that the error term is independent of the predictor variable.

While the true values of these parameters are unknown, they can be estimated using the least squares approach. Figure 3.5 shows a three-dimensional setting of linear regression model with two predictors and one response variable. In this figure, the difference between the observed response values and the values predicted by the regression model are represented as its vertical distance. By minimizing the sum of squared distances between each observation and the corresponding point on the fitted regression plane, the model determines the optimal plane that best captures the relationship between the two predictors and the response variable.

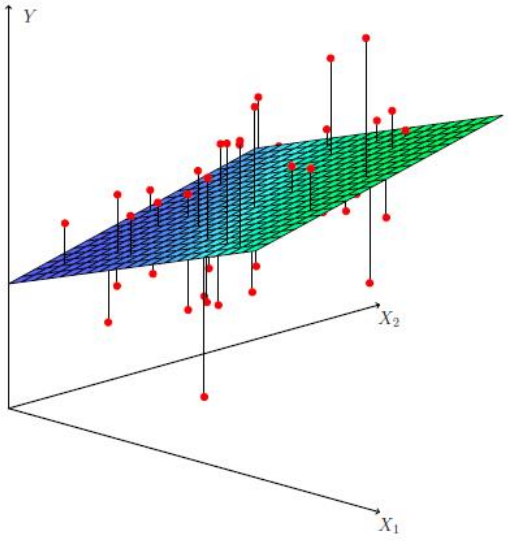


Figure 3.5 – Three-dimensional setting of linear regression model with two predictors and one response variable. The vertical lines present the difference between the observed response values and the values predicted by the regression model, and the plane represents the fitted regression model [14]

While the least square approach is useful for estimating the relationship between predictors and outcomes based on a specific dataset, it is essential to generalize this estimation to the entire population in order to obtain the true relationship between the variables.

On the left side of Figure 3.6, the true relationship between the predictor and outcome is represented by the red line, while the blue line corresponds to the least squares estimate based on a dataset. On the right side of the figure 3.6, the population regression line is displayed along with ten other estimates calculated using different sets of observations. Although each least squares line is different, on average, they closely approximate the population regression line. Therefore, getting an average of the estimates obtained from many datasets would accurately represent the true relationship.

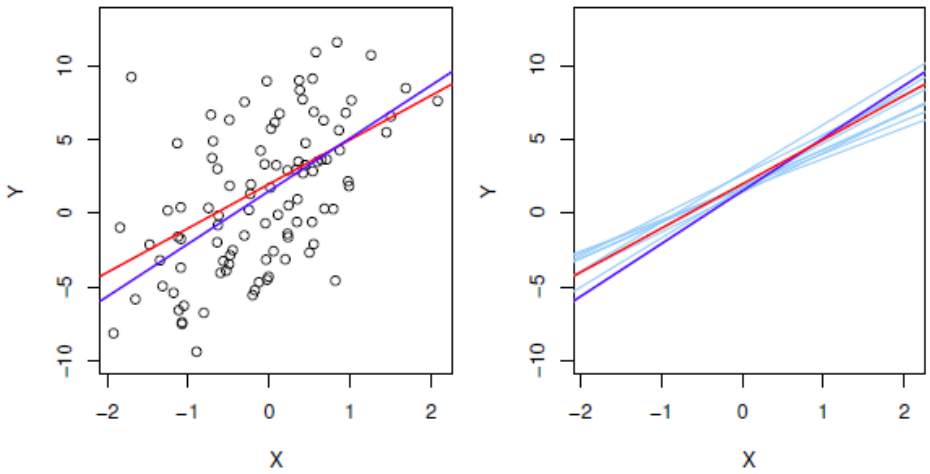


Figure 3.6 - The true relationship between the predictor and outcome is represented by the red line, while the blue line corresponds to the least squares estimate based on a dataset

Furthermore, in order to assess the results of the estimated coefficients, t-test can be performed on the coefficients. Choosing a confidence interval of 95%, a p-value smaller than 5% indicates that such a substantial association between the predictor and the response is unlikely to be observed due to chance, and there is a statistically meaningful association between the predictor and the response [14].

3.2.1 Evaluation metrics

After implementing a model, it is crucial to evaluate how well the model fits the data. One useful metric is the Mean Squared Error (MSE), which is obtained by the mean of the squared differences between each observed value (yi) and its corresponding predicted value (ŷi) (formula 8).

Formula 8 – Mean Squared Error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(yi - \hat{y}i)^2$$

To interpret the error metric in the original scale of the data, it is useful to take a square from MSE and provide the Root Mean Squared Error (RMSE) (formula 9).

Formula 9 - Root Mean Squared Error

$$RMSE = \sqrt{MSE}$$

Both MSE and RMSE are used to evaluate the performance of the model, with lower values indicating better model performance. However, the interpretation of the RMSE value is context-dependent and may vary across different domains or applications.

The R2 statistic offers an alternative measure of fit that addresses this issue. It represents the proportion of variance explained by the regression model and always takes on a value between 0 and 1, regardless of the scale of the outcome variable. While an R2 value close to 1 indicates that a large proportion of the variability in the response variable is accounted for by the regression model, an R2 value near 0 suggests that the regression model explains very little of the variability in the response [14].

3.2.2 Assessing the assumptions of linear regression

When using linear regression, there are some assumptions that are considered that need verification. The first basic assumption is the linearity, which assumes an existence of a linear relationship between the predictors and the response variable. Another crucial assumption is that the error terms are independent, have equal variance and are normally distributed.

It is important to note that the computation of standard errors for the estimated regression coefficients or fitted values relies on the assumption of uncorrelated errors. If there is correlation among the error terms, the estimated standard errors will tend to underestimate the true standard errors. Consequently, confidence and prediction intervals will be narrower than they should be. In time series data adjacent observations often have correlated errors. To investigate this, we can plot the residuals against time. If there is no noticeable pattern, the errors are likely uncorrelated. However, if adjacent residuals show tracking or similarity, it suggests error term correlation [14][25]. To assess the normality assumption, the histogram of the residuals can be plotted. If the residuals follow a roughly symmetric bell-shaped distribution, it suggests that the assumption of normality is met.

As mentioned in section 3.1.3, linear regression was used in the first part of this research as an attempt to model the regenerated values of heat pump power based on the most significant frequencies identified through Fourier transform analysis. Also, in the second part, it was used to

establish a relationship between the heat pump power and temperature. By subtracting the true values from the predicted values, the residuals were obtained. Fourier transform was then applied on the residuals to identify any remaining behavioural patterns, after eliminating the influence of temperature on the heat pump power.

4. Results and analysis

This section is composed of two parts:

In the first part, the Fourier transform was employed to calculate and analyse the recurring patterns in the heat pump power, and the temperature inside and outside of the house. Then, a linear regression was used to model the regenerated values of each parameter based on the most significant frequencies identified through Fourier transform analysis. The main objective is to which the heat pump power patterns are found in the temperature, and whether we can introduce a general pattern for the time series data, along with the varying factors specific to each house.

In the second part, a linear regression model was employed to model the overall behaviour of the heat pump power and establish a relationship between the heat pump power and temperature. This approach aimed to eliminate the influence of temperature on the heat pump power and identify any remaining behavioural patterns using Fourier transform.

4.1 Part 1

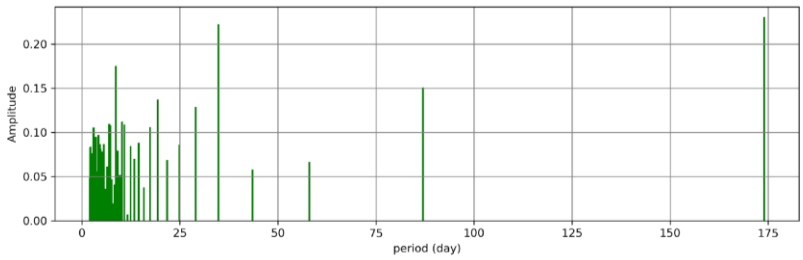
4.1.1 Fourier transform on original values

By applying the Fourier analysis of the normalized daily power usage, as well as the indoor and outdoor temperatures for each house, the underlying patterns within the data were identified. Figure 4.1 shows and example of such an implementation for an example house, where the amplitude and period of the patterns found in each house is plotted. In this figure, the x-axis displays the periods, which represent the durations of periodic patterns present in the data. The y-axis represents the amplitude, which signifies the strength or magnitude of each periodic component.

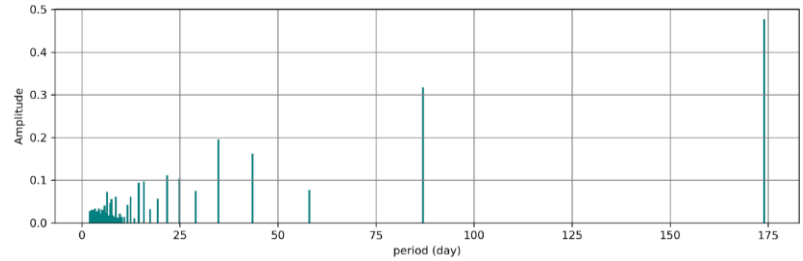
It can be seen that the highest amplitude corresponds to the period of 175 days, which aligns with the entire duration of the data. This period is associated with a frequency near zero in the Fourier analysis, which as discussed in section 3.1.2, represents the average value of the signal over the entire time period, and does not contribute to any oscillatory patterns. Also, the accumulation of low-amplitude periods within the range of 0 to 25 days makes it less efficient to distinguish meaningful patterns from noise. To identify the significant frequencies in the data, the top ten peaks were calculated for each variable and each house.

Figure 4.2 provides an overview of the top ten frequencies observed across all houses. Figure 4.2-a and Figure 4.2-b shows the amplitude (on the y-axis) and periods (on the y-axis) respectively, focusing on the prominent repeating patterns discovered in the normalized heat pump power and temperature within the houses. Additionally, Figure 4.2-c visualizes the distribution of periods detected in the temperature outside the houses through a count plot. The y-axis of this plot represents the number of occurrences of each period for all houses, for example: a pattern with a

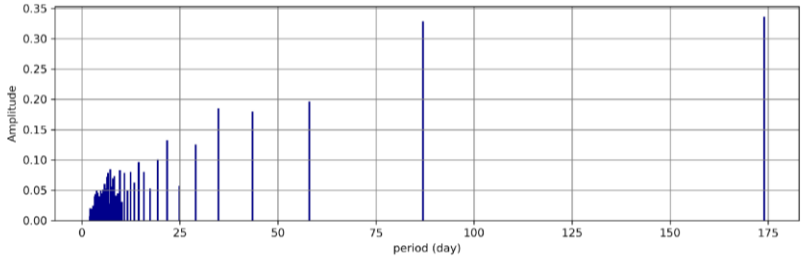
period of 11 days was found in the temperature outside of 40 houses. This indicates that every 11 days, there was a recurrent change in temperature observed in these houses.



(a) Normalized heat pump power



(b) Normalized temperature inside the house



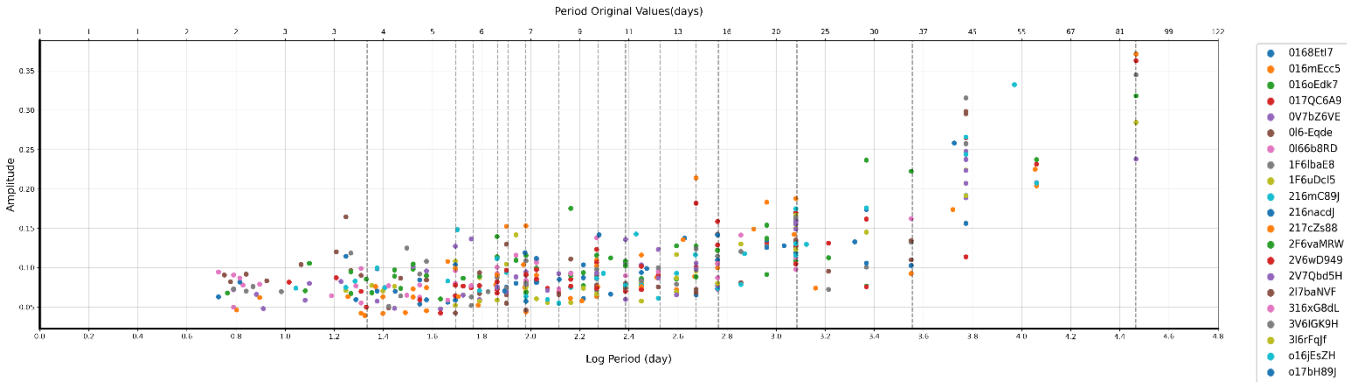
(c) Normalized temperature outside the house

Figure 4.1 - The corresponding periods obtained from Fourier transform for house id *z17bDdY4*

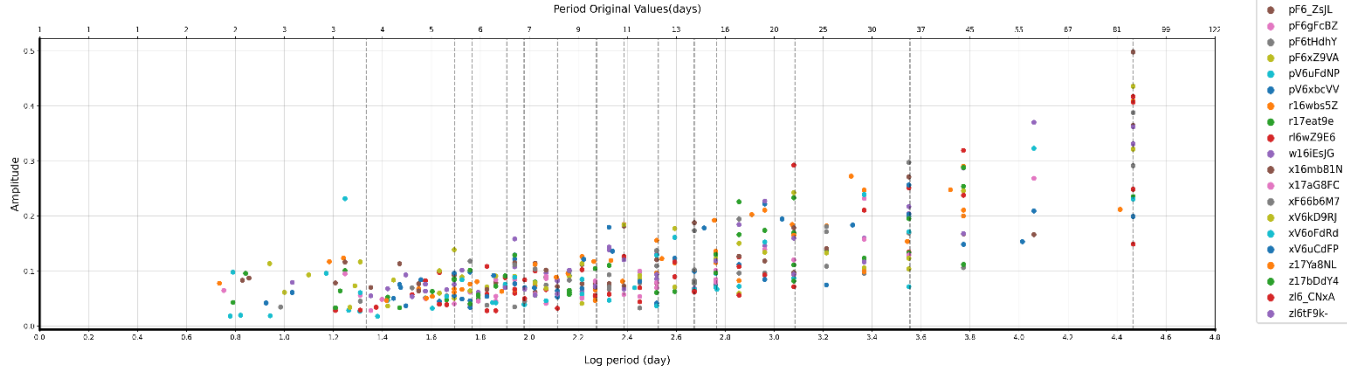
By examining periods with a count exceeding 10, it becomes evident that these patterns are also noticeable in both heat pump power usage and temperature inside the houses. These observed periods are marked with dashed lines, indicating the occurrence of these patterns in all three figures. This correspondence is not unexpected, since heat pump power usage is influenced by the outside weather for each house. What is more interesting to see is the shared patterns in the heat pump power usage and temperature inside the houses, which are not found in the temperature outside of the house. This can suggest the presence of behavioural patterns in heat pump power usage that are independent of the external temperature.

However, simply identifying patterns is not enough; it is crucial to have a solid explanation for why these patterns occur. To make use of this information effectively, we need a strong theory that explains the reasons behind these patterns. For example, let's consider the interesting 6-7 day period, which coincides with weekends. It could indicate a societal behaviour where people spend

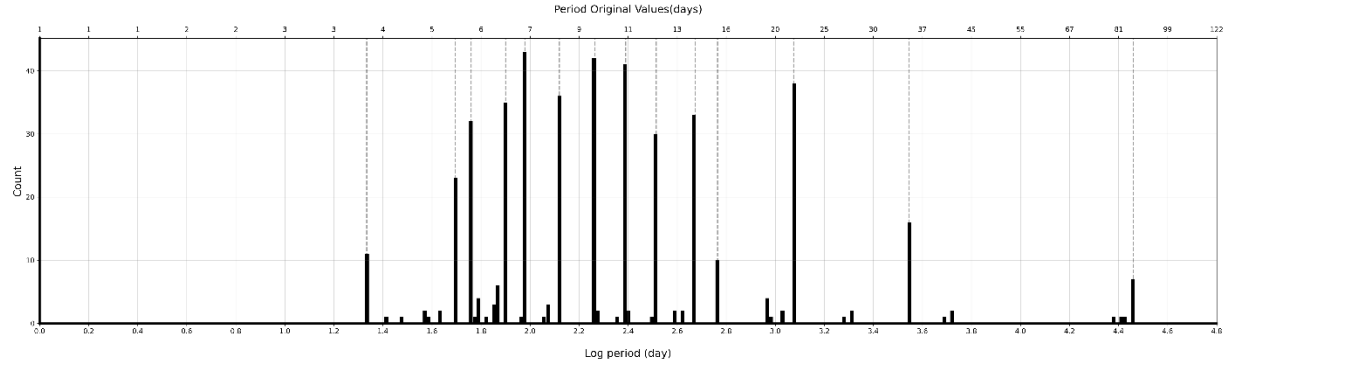
less time at home, leading to reduced heat pump power usage. However, the data does not clearly indicate whether this pattern is solely due to people's behaviour or if it is primarily driven by changes in weather. Both factors are present in the data, and it remains uncertain which one is responsible, or if both factors contribute to the observed pattern.



(a) Fourier transform results of heat pump power



(b) Fourier transform results of temperature inside of the house



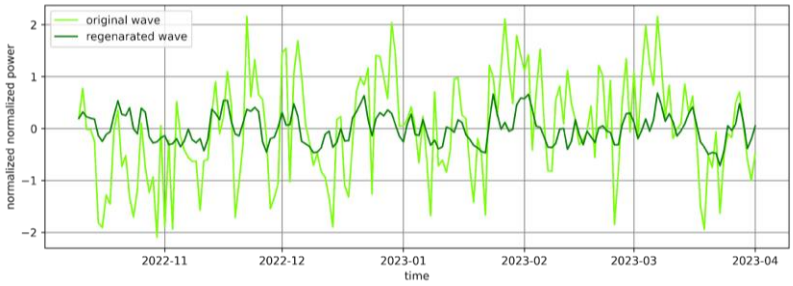
(c) Count plot of Fourier transform results of temperature outside of the house

Figure 4.2 - Fourier transform results

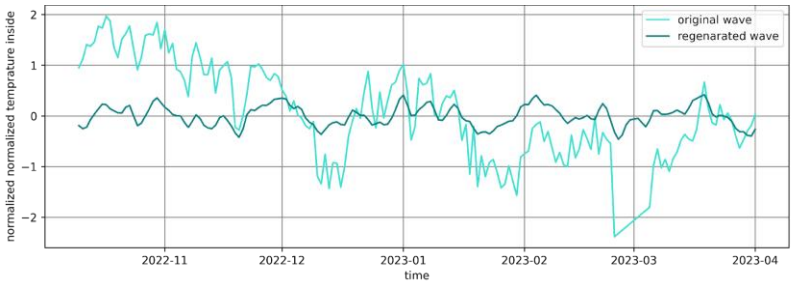
4.1.2 Linear regression on the regenerated values

A linear regression analysis was performed for all houses to investigate the relationship between the ten most important patterns found in the heat pump power and the temperatures inside and outside of the house. The regression model included the regenerated signal from the top ten frequencies for each variable. The objective was to investigate the extent to which the frequencies identified in the temperatures can explain the frequencies observed in the heat pump power.

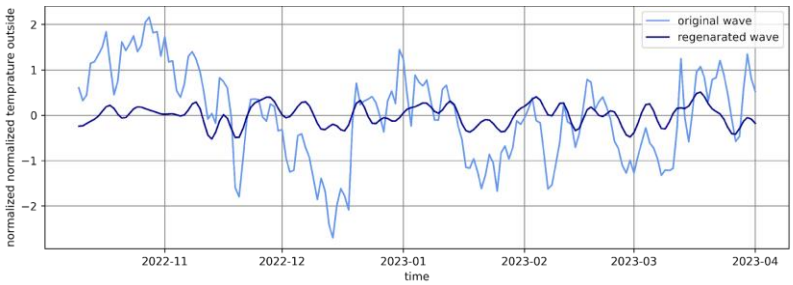
Figure 4.3 provides a visual representation of the reconstructed signal for each variable during time for an example house. It is important to note that while the regenerated signal captures the main fluctuations and patterns, it may not exactly match the original signal due to the omission of several Fourier terms.



(a) Normalized heat pump power



(b) Normalized temperature inside the house



(c) Normalized temperature outside the house

Figure 4.3 – the regenerated signal from ten top peak frequencies in the data obtained from Fourier transform for house id 'z17bDdY4'

Once the signals were constructed, they were used as input for a multiple linear regression model. The regenerated signal obtained from the top ten peak frequencies of the normalized heat pump power served as the outcome variable, while the temperatures inside and outside of the house were utilized as predictor variables.

To evaluate the effectiveness of the model, a random division of house IDs was performed, with 2/3 of the houses allocated for training (30 houses) and 1/3 for testing (15 houses). To assess the model's performance, R2 value and mean squared error were calculated.

To obtain a more accurate estimation, this process was repeated 1000 times, and the mean value of the evaluation metrics were calculated. The results showed a RMSE of 0.26, and R2 value of 5.38%. [1]Considering that the variables were z-normalized, the RMSE of 0.26 signifies that on average, the model's predictions deviate by 0.26 standard deviations from the actual values of the response variable. Furthermore, the resulted R2 value indicates that the regression model explains only a small portion 5.38% of the variability in the response variable. This low R2 value suggests that the predictor variables may not be effectively capturing the underlying factors that influence the response variable. Therefore, the model may not be a good fit for the data, and alternative approaches may need to be considered to improve the model's performance and increase the proportion of variance explained. Another approach is employed in part 2, which will be discussed in section 4.2.

These calculations were repeated using 25 top peaks instead of 10, but the result did not improve. The reason 25 top peaks were chosen was they were the optimal value to reconstruct the values in a way that it is very close to the original values. More information about this can be found in Figure B-1 Appendix B.

To visualize and interpret the model's coefficients and predicted values, a specific train-test combination was selected that had an R2 value close to the mean 5.38% . The Table B-1 in Appendix B provides the house IDs for the test and train sets. The coefficients of the model are 0.04 and -0.29 and the intercept is nearly zero. The significance of the coefficients in the linear regression model was assessed using a 95% confidence interval. The t-statistics and p-values of the model are presented in Table 4.1.

Table 4.1 - T-statistics and p-values of the created model

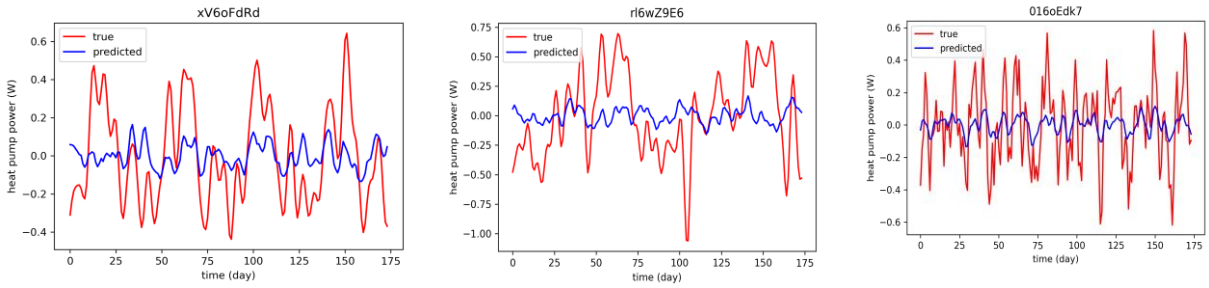
	coef	std err	t	P>|t|	[0.025	0.975]
const	2.198e-17	0.004	6.21e-15	1.000	-0.007	0.007
t_in f	0.0437	0.012	3.706	0.000	0.021	0.067
t_out f	-0.2911	0.016	-18.501	0.000	-0.322	-0.260

On the other hand, the p-value for both the normalized temperature outside and inside variable is smaller than 5%. This indicates that there is a significant relationship between the normalized

1. The full results can be seen in figure B-2 of Appendix B

power and the normalized temperature outside and inside of the house. Specifically, for a given amount of normalized temperature inside the house, an increase in the normalized temperature outside the house is associated with a 0.29 unit decrease in normalized power. This aligns with the expected reality that as the temperature outside the house increases, more power is required to maintain a comfortable temperature inside.

In Figure 4.4, shows some[1] of the predicted values generated by the model are represented by the blue line, while the true values are depicted in red. The predicted signal captures some patterns available in the data. However, it is important to note that the predicted values do not align closely with most of the original signal values.



(a) (b) (c)

Figure 4.4 - predicted values for the normalized power in the test set, versus the true values

By plotting the residuals (figure 4.5), it can be seen that the majority of the residuals cluster around the mean, and the distribution of residuals appears to be normal, with a slight left skew observed for certain house IDs. A slight left skewness indicates that some of the residuals have lower values than expected based on a normal distribution. However, the deviation is not considerable, and overall, the assumption of normality can be reasonably justified.

To check the randomness in the data, the autocorrelations were plotted at varying time lags. Figure 4.6 shows the autocorrelation coefficients on the vertical axis and the time interval on the horizontal axis. Furthermore, the horizontal dashed lines represent the 95% confidence interval for statistical significance. The autocorrelation function evaluates the correlation between observations in a time series across different lags. This is why at lag 0, the autocorrelation coefficients are equal to 1, as they are representing the correlation of the feature with itself at the same time point.

It can be seen that different devices show various patterns. Some devices show autocorrelation coefficients that remain within the dashed lines, indicating that their residuals are not significantly correlated. House id *016oEdk7* shows such a pattern, and if this is compared with figure 4.4 – c, it can be seen that the predicted values captured the main fluctuations of the timeseries.

However, some devices such as *xV6oFdRd* show a clear recurring patterns every 20 days, while house id *rl6wZ9E6* exhibits such a pattern every 45 days. As mentioned before in section 3.2.2,

the presence of recurring patterns or significant autocorrelation in the residuals suggests that the model did not fully capture the patterns available in the data, and this can be clearly seen in figure 4.4-a and figure 4.4-b.

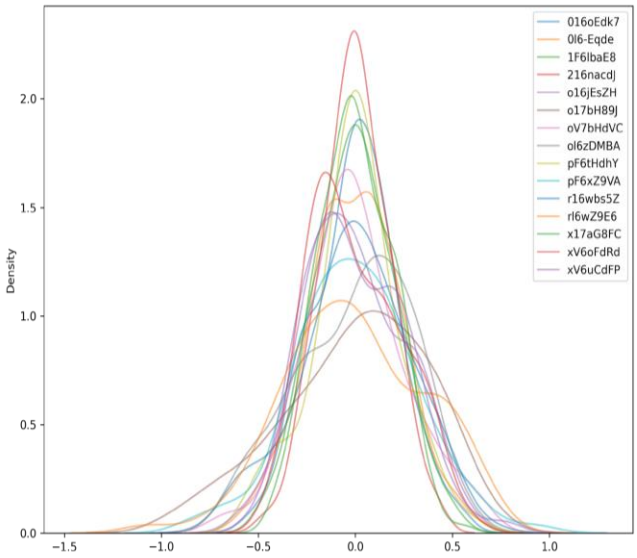


Figure 4.5 - A plot of the residuals to assess the assumption the normality of residuals

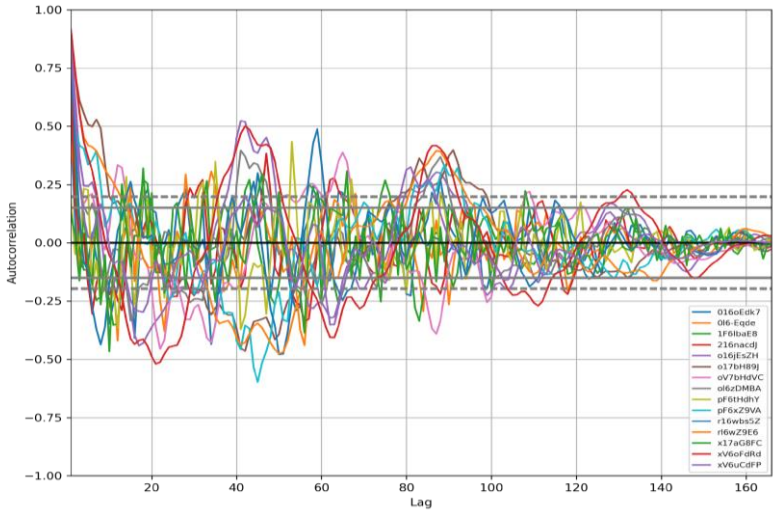


Figure 4.5 - A plot of the autocorrelation to assess the assumption uncorrelated errors

4.2 Part 2

4.2.1 Linear regression on original values

In section 4.1, the performance of the multiple linear regression model based on the regenerated waves from the top peaks in the Fourier transform was unsatisfactory. Consequently, an alternative approach was adopted, involving a linear regression on the normalized heat pump power along with temperature measurements inside and outside of the house.

Similar to the previous section, a random partitioning of house IDs was carried out, assigning 2/3 of the houses for training purposes and 1/3 for testing. This process was repeated 1000 times using different seeds, and the average values of the evaluation metrics were computed. The resulting mean Root Mean Squared Error (RMSE) was 0.72, indicating that the model's predictions deviated by 0.72 standard deviations from the actual values of the response variable. The mean R2 value increased to 46.96, indicating a significant improvement in the model's performance compared to the initial approach.

Further investigation and in-depth analysis of this model are beyond the scope of this thesis. If the reader is seeking a more comprehensive understanding of the specifications of the linear regression method, they may refer to another thesis topic within this series, Finding human behavioural in heat pump power usage using predictive modelling by Ruben de Groot.
The ultimate finding from these linear regression models is that the temperature proves to be a strong indicator of heat pump power usage. This finding aligns with the real-world observations: it is only logical that people consume more power during cold weather and when they desire to keep their homes warmer. However, there is still a crucial question left unanswered: Can we spot the effects of human and societal behavioural patterns in the data, after accounting for temperature? This interesting aspect deserves more investigation in future research.

4.2.2 Residuals of linear regression model

The residuals obtained from the model have the potential to uncover the societal behavioural patterns we are searching for. These residuals represent the portion of the data that cannot be explained by temperature alone. By subtracting the predicted values from the actual ones, we are left with the part that is not accounted for. However, it is important to note that the low number of observations specially in the second part of this study limits the statistical certainty of these findings.

Figure 4.7 shows the plot of residuals for 10 houses in the test set. Although the residuals mainly gather around the zero baseline, there are noticeable deviations observed in certain dates. In other words, the model is predicting a different power usage compared to the actual values. These deviations are unexpected, given the fact that heat pump power usage usually follows the outside

temperature closely. To investigate the reasons behind these deviations, the average outside temperature for all houses was calculated. One plausible hypothesis is that deviations from the expected pattern of heat pump power usage occur on days characterized by substantial temperature changes. Figure 4.8 shows a plot of the temperature outside of the house and Table 4.2 presents the dates where the mean temperature of a particular day shows a difference exceeding 5 degrees in comparison to the previous day. To some extent, these dates align with the peculiar patterns observed in the residuals.

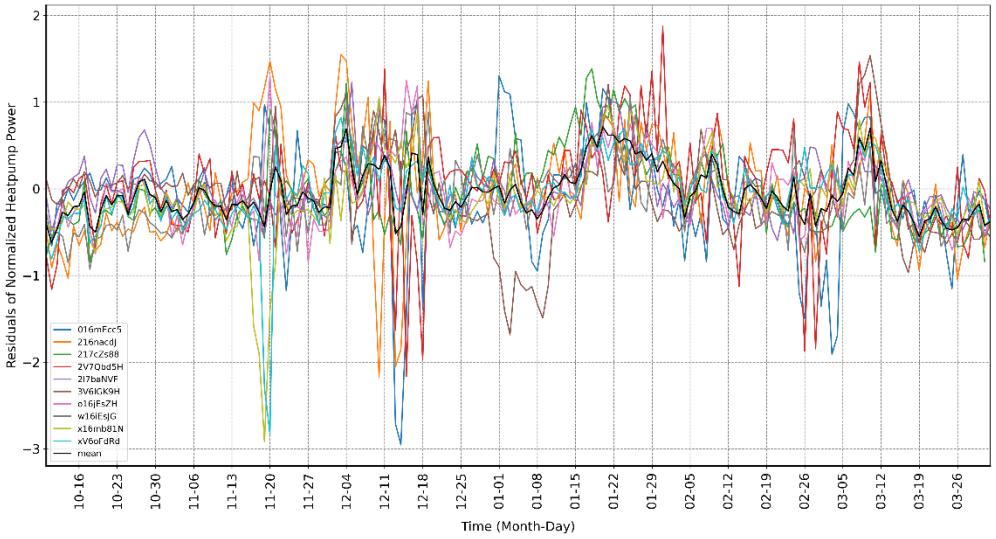


Figure 4.7 – The plot of residuals for 10 houses in the test set

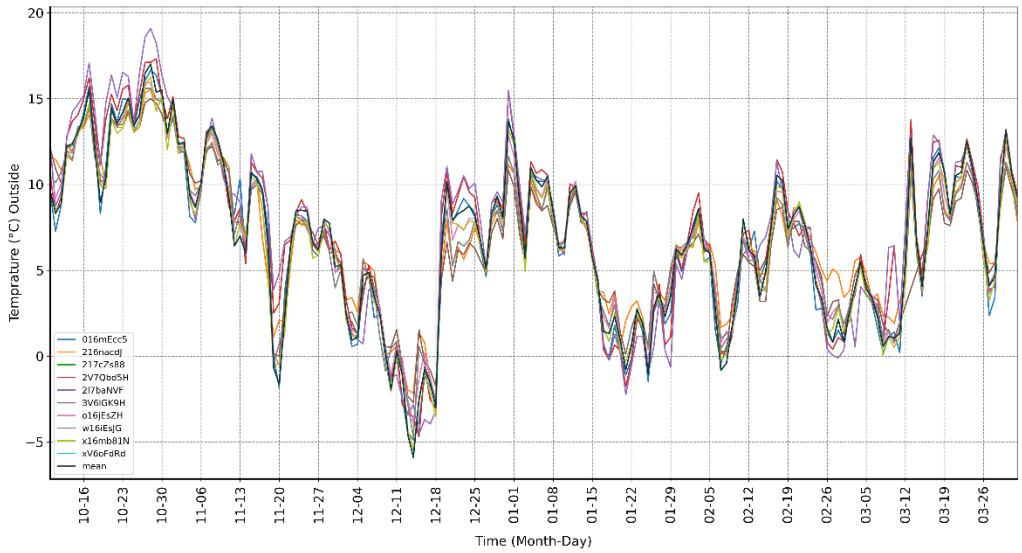


Figure 4.8 – The plot of temperature outside of the house for 10 houses in the test set

Table 4.2 - dates where the mean temperature of a particular day shows a difference exceeding 5 degrees in comparison to the previous day

Date	Temperature difference (°C)
2022-11-19	-7
2022-12-19	+9
2022-12-31	+6
2022-01-04	+5
2023-03-13	+7
2023-03-14	-6

One possible explanation could be that people tend to adjust their heat pump settings with a certain delay, typically when the indoor temperature deviates from their desired comfort level over a period of time. Consequently, when there is a sudden weather change compared to the previous day, the heat pump power usage does not instantaneously adapt and align with the outside temperature. This delay in response becomes visible in the model residuals, resulting in the observed deviations.

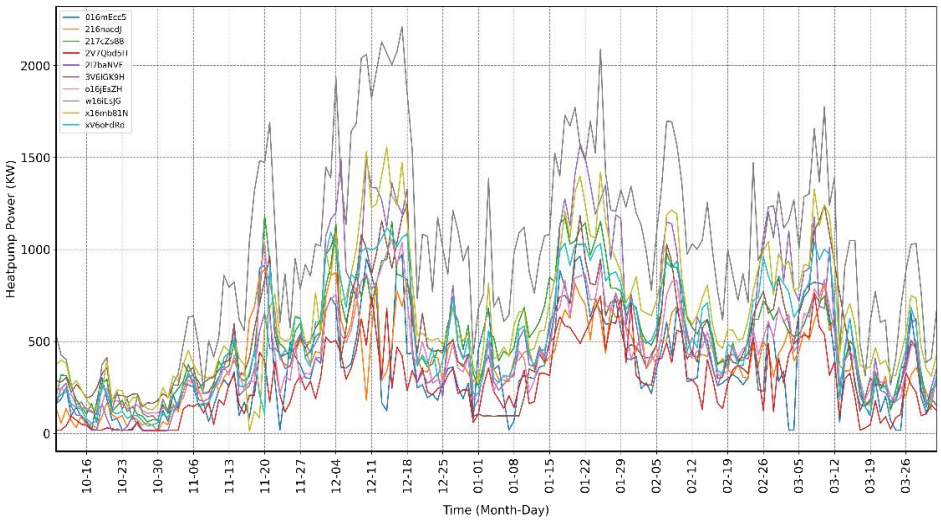


Figure 4.9 – The plot of heat pump power usage of the house for 10 houses in the test set

While this theory already suggests a potential behavioural pattern, a clearer indication of societal patterns emerges when looking at the residual plots of house IDs *3V6IGK9H* and *016mEcc5*. These houses show lower power usage during the period from December 1st to 8th and during the week of February 26th to March 5th, respectively, and these dates align with the Christmas holiday and the spring holiday for schools in different provinces (figure 4.9). These deviations cannot be explained by the previous hypothesis, but they can be justified by considering that the occupants

of these houses were on vacation during those periods. It is plausible that they intentionally minimized their energy usage by turning off their heat pump while they were away. To gain more comprehensive insights into this matter, further investigation is necessary, including the analysis of the data from additional houses and considering holidays in April and May, which are more than those in the investigated period.

4.2.3 Fourier transform on residuals

Similar to the previous section, the residuals were subjected to the FFT algorithm, aiming to extract the most important frequency components hidden within the time series dataset. In Figure 4.10, the logarithm and the real value of the pattern period are present in the residuals for each house are shown in the lower and upper horizontal axis respectively, by their corresponding amplitudes obtained from the Fourier transform on the y axis.

By comparing the Fourier results showing patterns in the heat pump power (Figure 4.10) and the corresponding residuals for the same 10 houses (Figure 4.11), several observations can be made. The original heat pump power patterns associated with 6 days, 9-11 days, 11-13 days, 16 days, and 20-25 days are largely diminished in the residuals. However, certain houses still exhibit residual values that demonstrate recurrent patterns at approximately 6-7, 13, 25, and 45 days. This suggests that even after removing the influence of temperature on the heat pump power, there are still periodic changes in power usage that are visible on the Fourier analysis of residuals. These recurring patterns may indicate behavioural factors, such as weekly or biweekly activities that cause people to be away from their homes, resulting in changes in power usage.

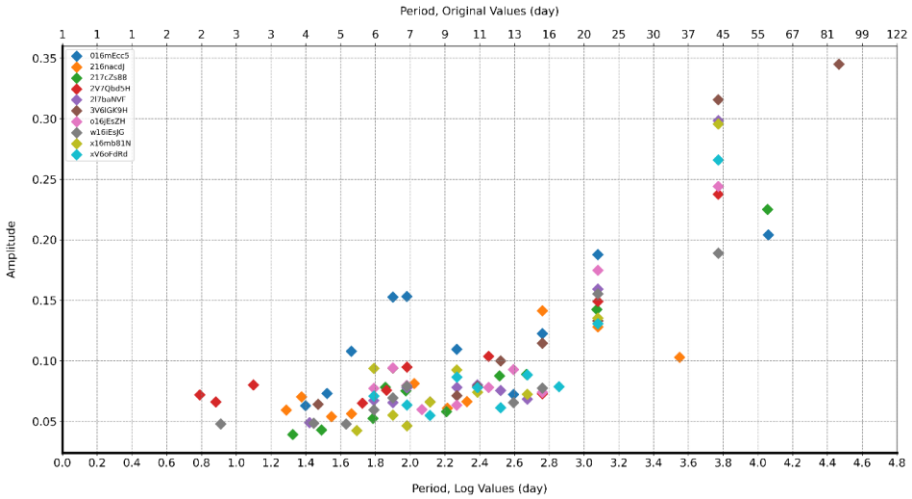


Figure 4.10 - The logarithm and the real value of the pattern period are present in the original values of heat pump power

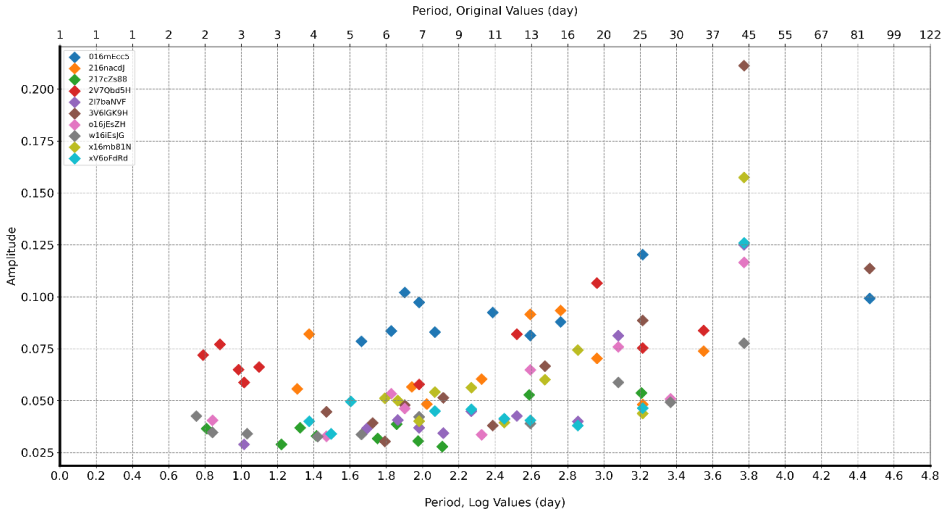


Figure 4.11 - The logarithm and the real value of the pattern period are present in the residuals of heat pump power

5. Conclusions

The findings of this study do not provide conclusive evidence of identifying human behavioural patterns in relation to heat pump usage. While few patterns were observed, the limitations in the number of identified patterns and available measurements prevent us from confidently attributing them to human behaviour. Further research and a larger sample size may be necessary to gain a more comprehensive understanding of the relationship between human behaviour and heat pump usage.

The results of this study, in conjunction with the linked study by Ruben de Groot, indicate that temperature outside the house plays a significant role in influencing heat pump power usage. The application of Fourier analysis in the first part successfully revealed patterns that are present in both temperature and heat pump power usage, as well as patterns that are unique to the temperature inside the house and heat pump power alone. This was somehow confirmed in the second part. Some possible patterns that are related to the holidays or weekends could be spotted, by analysing the residuals of the linear regression model and implementing Fourier analysis on them. However, it is important to note that the low number of observations specially in the second part of this study limits the statistical certainty of these findings. Further research with a larger sample size would be beneficial in confirming these patterns and their relationship to heat pump power usage. Some suggestions for future research include expanding the analysis to a wider period that includes holidays such as the Easter period, May (when there are more national holidays in the Netherlands), and the summer period. Also, performing Fourier analysis on the data that is aggregated hourly would provide insights into the usage patterns and fluctuations of heat pump power throughout the day.

The results of linear regression on the regenerated values from the top frequencies were unsatisfactory, suggesting that this method may not be effective. The reasons for this are unclear, and further investigation is required. Furthermore, alternative statistical approaches, such as clustering techniques, could be explored to group houses based on the results of the Fourier transform analysis and uncover meaningful patterns within the data.

In addition to the methods used in this research, there are opportunities for improvement in the data preparation process as well. Specifically, the imputation method used, linear imputation, could be enhanced by considering other techniques such as Multiple Imputation by Chained Equations (MICE), which has demonstrated success in handling missing data. Furthermore, the current outlier detection method did not account for low temperatures, leading to outlier peaks, and it did not detect outliers for devices with missing values. Exploring alternative outlier detection techniques would be beneficial in addressing these limitations and ensuring more comprehensive data quality assessment.

The limitations of this research were primarily related to the time constraints and the availability and quality of data. The study was conducted within a relatively short time period of two months,

which limited the extent of analysis. Additionally, the project encountered server issues at the beginning, resulting in a loss of several weeks of data. These time limitations may have affected the depth and breadth of the research findings. Furthermore, the data itself was noisy, incomplete and full of missing data, which could have introduced uncertainties and impacted the accuracy of the results. It is important to acknowledge these limitations when interpreting the findings of this study.

Bibliography

[1] Holli Riebeek, "Global Warming," *NASA Earth Observatory*, Jun. 03, 2010. https://earthobservatory.nasa.gov/features/GlobalWarming

[2] "IPCC report: 'Code red' for human driven global heating, warns UN chief," *UN News Global Perspective Human Stories*, Aug. 09, 2021. https://news.un.org/en/story/2021/08/1097362

[3] United Nations, "The Paris Agreement," *United Nations, Climate Action*, Nov. 2022. https://www.un.org/en/climatechange/paris-agreement

[4] "The Netherlands 2020 Energy Policy Review," *IEA*, Sep. 2022. https://www.iea.org/reports/the-netherlands-2020

[5] Ministerie van Economische Zaken en Klimaat, "Klimaatakkoord," *Klimaatakkoord*, Jun. 28, 2019. https://www.klimaatakkoord.nl/klimaatakkoord/documenten/publicaties/2019/06/28/klimaatakkoord

[6] Ministerie van Economische Zaken, Landbouw en Innovatie, "Wat is het Klimaatakkoord?," *Rijksoverheid.nl*, Jan. 22, 2020. https://www.rijksoverheid.nl/onderwerpen/klimaatverandering/klimaatakkoord/wat-is-het-klimaatakkoord

[7] Ministerie van Algemene Zaken, "Measures to reduce greenhouse gas emissions," *Government.nl*, Jun. 02, 2021. https://www.government.nl/topics/climate-change/national-measures

[8] "Hybride warmtepompen, haalbaar en betaalbaar," *Nederlandse Verduurzamings Industrie - Gebouwde Omgeving*, Nov. 21, 2021. https://www.nvi-go.nl/publicaties/hybride-warmtepompen-haalbaar-en-betaalbaar/

[9] Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, "Warmtepomp de norm vanaf 2026: goed voor klimaat en de energierekening," *Rijksoverheid.nl*, May 02, 2023. https://www.rijksoverheid.nl/actueel/nieuws/2023/05/01/warmtepomp-de-norm-vanaf-2026-goed-voor-klimaat-en-de-energierekening

[10] "Demonstratie- project Hybride warmtepompen in de gebouwde omgeving," *Demoprojecthybride*. https://www.demoprojecthybride.nl/ (accessed Jul. 08, 2023).

[11] Milieu Centraal, "Warmtepomp: duurzaam elektrisch verwarmen," *Milieu Centraal*. https://www.milieucentraal.nl/energie-besparen/duurzaam-verwarmen-en-koelen/warmtepomp-duurzaam-elektrisch-verwarmen/ (accessed Jul. 08, 2023).

[12] "Eindrapportage Installatiemonitor 1.0 - Installatiemonitor.nl," *Installatiemonitor.nl*, Feb. 21, 2023. https://www.installatiemonitor.nl/eindrapportage-installatiemonitor-2/

[13] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Elsevier, 2011.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media, 2013.

[15] S. Van Buuren, *Flexible Imputation of Missing Data, Second Edition*. CRC Press, 2018.

[16] P. K. Janert, "Data Analysis with Open Source Tools," *O'Reilly Online Learning*. https://www.oreilly.com/library/view/data-analysis-with/9781449389802/ch04.html

[17] "Prof. Paul Cuff ELE 201: Information Signals Course Notes," season-01 2016. https://www.princeton.edu/~cuff/ele201/

[18] Wikipedia, "Time series," *Wikipedia*, Jun. 2023, [Online]. Available: https://en.wikipedia.org/wiki/Time_series

[19] Libretexts, "2.3: Signal Decomposition," *Engineering LibreTexts*, May 2022, [Online]. Available: https://eng.libretexts.org/Bookshelves/Electrical_Engineering/Introductory_Electrical_Engineering/Electrical_Engineering_(Johnson)/02%3A__Signals_and_Systems/2.03%3A_Signal_Decomposition

[20] Wikipedia, "Fourier transform," *Wikipedia*, Jun. 2023, [Online]. Available: https://en.wikipedia.org/wiki/Fourier_transform

[21] Q. Kong, T. Siauw, and A. Bayen, *Python Programming and Numerical Methods: A Guide for Engineers and Scientists*. Academic Press, 2020. [Online]. Available: https://pythonnumericalmethods.berkeley.edu/notebooks/chapter24.01-The-Basics-of-waves.html

[22] D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics*. John Wiley & Sons, 2013.

[23] J.-B. Yao, B. Tang, and J. Zhao, "Improved discrete Fourier transform algorithm for harmonic analysis of rotor system," *Measurement*, Apr. 2016, doi: 10.1016/j.measurement.2016.01.028.

[24] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*. Princeton University Press, 2003.

[25] J. J. Faraway, *Linear Models with R*. 2004. doi: 10.4324/9780203507278.

Appendix A

The time resolution column in table A-1 indicates the frequency at which new measurements were recorded for each sensor. However, as these time resolutions were more detailed than necessary for our analysis, the data was aggregated on a daily basis. The aggregation techniques for each parameter are indicated in the column Aggregation.

Table A-1 - A summary of the sensors and measurements in the Demo Project Hybrid

General	Quantity	Min	Max	Unit	Time resolution [s]	Aggregation
Heating system sensor (Kampstrup)	**T_supply:** System supply temperature	-50	150	degC	5	Mean
	T_return: System return temperature	-50	150	degC	5	Mean
	E_thermal_system: kWh heat sum	-1000	1000000	kWh	5	Cum (max)
	P_thermal_system: kW heat instantaneous (power)	-100000	100000	W	5	Mean
	F_system: Water flow running throught the system	-100	100	liter/min	5	Mean
Heatpump sensor	**Energy:** kWh electricity sum heat pump	0	1000000	kWh	60	Cum (max)
	Power: (W electricity at a point heat pump)	0	10000	W	5	Mean
Boiler sensor	**Energy:** kWh energy sum boiler	0	1000000	kWh	60	Cum (max)
	Power: amount of W at a point for boiler	0	10000	W	5	Mean
Smart meter (DSMR)	**p_consumed_L1:** kW_laag fase 1 (kW low phase 1)	-50000	50000	W	10	Mean
	P_consumed_L2: kW_laag fase 2	-50000	50000	W	10	Mean
	P_consumed_L3: kW_laag fase 3	-50000	50000	W	10	Mean
	P_delivered_L1: kW_hoog fase 1 (kW high phase 1)	-50000	50000	W	10	Mean
	P_deliverd_L2: kW_hoog fase 2	-50000	50000	W	10	Mean

	P_delivered_L3: kW_hoog fase 3	-50000	50000	W	10	Mean

Appendix B

In order to determine the optimal number of top frequencies that capture the key variations in the data, a range of different peak numbers, from 1 to 84, were tested. For each number of peaks, the error between the original data and the regenerated waves using those peaks was calculated. The results are presented in the plot below, clearly demonstrating that beyond 25 peaks, the error showed minimal change. This suggests that utilizing 25 peaks is a favourable choice, which was consequently employed in the regression model. However, increasing the number of peaks did not result in any noticeable improvement in the results. As a result, the decision was made to retain 10 peak frequencies, as they were deemed sufficient for capturing the essential characteristics of the data.

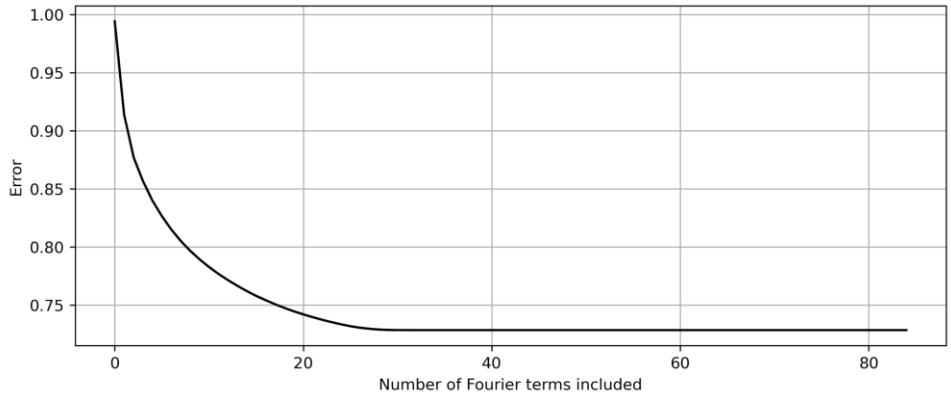


Figure B-1 – Plot of the errors based on the number of peaks involved in the regenerated values

Table B-1 – the trains and test set ids

seed	225
Train set ids	'0168EtI7', '016oEdk7', '017QC6A9', '0l6-Eqde' ,'0l66b8RD' ,'1F6lbaE8', '1F6uDcI5', '216mC89J' ,'216nacdJ' ,'2F6vaMRW', '316xG8dL', 'o16jEsZH', 'o17bH89J' ,'oV7bHdVC', 'ol6zDMBA' ,'pF6_ZsJL', 'pF6tHdhY', 'pF6xZ9VA', 'pV6xbcVV' ,'r16wbs5Z' ,'r17eat9e' ,'w16iEsJG' ,'x16mb81N' ,'x17aG8FC', 'xF66b6M7' ,'xV6uCdFP' ,'z17Ya8NL', 'z17bDdY4', 'zl6_CNxA', 'zl6tF9k-'
Test set ids	'016mEcc5', '0V7bZ6VE', '217cZs88' ,'2V6wD949', '2V7Qbd5H' ,'2l7baNVF', '3V6lGK9H' ,'3l6rFqJf' ,'o17dHNRB' ,'oV7cDMBO', 'pF6gFcBZ', 'pV6uFdNP', 'rl6wZ9E6' ,'xV6kD9RJ' ,'xV6oFdRd'

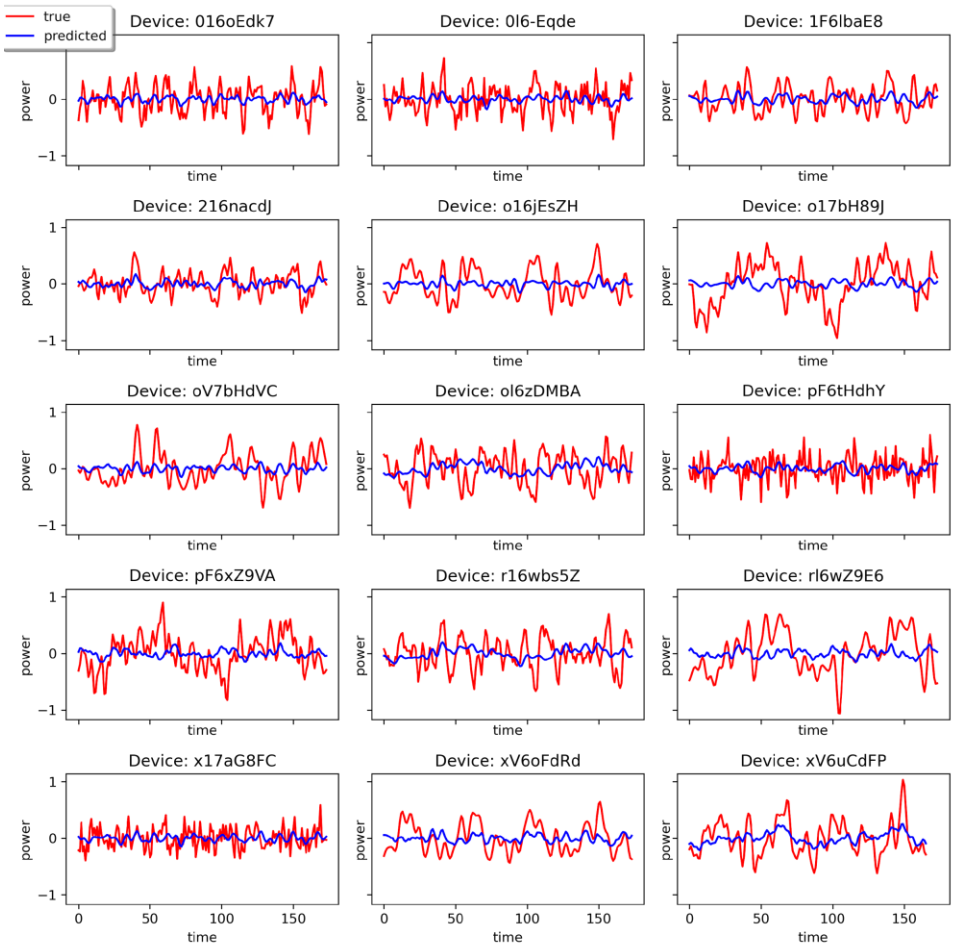


Figure B-2 – Plot of the predicted values based on the model and the original values the heat pump power