Building a Predictor for Electricity Consumption per Household Type

Alexandru-Ioan Chelba

MInf Project (Part 1) Report

Master of Informatics
School of Informatics
University of Edinburgh

2022

Abstract

This project investigates the problem of predicting electricity consumption per household type. The analysis is relevant in the context of moving from analysis per individual household to analysis per subsector level inside the residential sector. Given that renewable energy needs to be stored in order to be accessible at any given time of the day, knowledge of approximating the amount of electricity needed in the market is becoming paramount to optimizing costs and storage capacity. Much of the work to date in the domestic sector has focused on the individual household and improving prediction accuracy for one household. This project uses data which looks specifically at single-family homes in and around Edinburgh. Using this data, the homes were grouped into categories and data for each category was aggregated in order to obtain representative datasets for all categories. This paper shows the analysis done for three of the categories. It then constructs a predictive model based on the analysis, with parameters customised to fit each dataset in the best way. Results show that the model outperformed the baseline, proving that the exploratory analysis offers good grounding for understanding the patterns in the data.

Acknowledgements

I would like to thank my supervisor, Nigel Goddard, for the dedicated time, all pieces of advice and the web of support he provided me with throughout the year. I would also like to thank my family for understanding my needs and providing me with the necessary energy to get this done.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Global climate change has been an issue that has gained wide recognition over the years [32] [31]. The vast amount of carbon dioxide emissions has been recognised as an important factor in global warming and has been gaining increased attention from policymakers, especially since electricity and heat production accounts for 25% of the emissions [14]. The Scottish Government's current target is for Scotland to become carbon-neutral by 2045, while its capital, Edinburgh, attempts to achieve carbon-neutrality by 2030, according to Edinburgh City Council [35].

To tackle the massive challenges set by this goal, new forms of renewable energy are becoming prevalent in the market, such as solar and wind. The main issue with these is that the sources are not available non-stop, so energy cannot just be generated as demand increases. Another issue is that electricity storage is limited [20], so it is essential that supply matches demand. Additionally, turning on the power station is not easy, so it must be well planned.

In order to find a solution, accurately predicting demand is of utmost importance. This has been done for big industrial companies for a long time, but never with homes, due to individual demand being low. However, in total, the domestic sector accounts for 36% of the total energy consumption in the UK [23], a significant share of the market. Effective household energy management thus becomes crucial in reducing the carbon footprint overall.

In the UK, demand side response (DSR) has been developed as a tool to ensure a secure, sustainable and affordable electricity system [29]. Energy providers give financial incentives to users to turn down or turn off non-essential processes at times of peak demand, helping the grid balance supply and demand without the need for additional generation (e.g. power stations). New technology may further help with this. For example, an electric vehicle charger lets the user choose how much of the battery they want to charge, and until when they want this to happen.

As part of a bigger research project which attempts to address the problems described above [1], this paper seeks to build an understanding of electricity consumption per

type of house and create a predictor that is able to take input parameters and output next-day electricity demand, averaged for every half-hour. This aligns with industry standards, where predicting at network level is preferred to predicting at individual level and where 30-minute measurements are common, with smart meters installed in homes often reporting data at this frequency.

To accomplish the task, the paper makes use of the IDEAL (Intelligence Domestic Energy Advice Loop) dataset, which is accessible at [2] with the original documentation available at [28]. The data was collected by researchers from the University of Edinburgh and contains both electricity and gas sensor data, as well as various metadata and surveys collected from 255 households over 20 months between 2016 and 2018 in and around Edinburgh.

1.2 Objectives

The main goals of the project for this year are:

- Produce a model capable of detecting patterns in electricity consumption per type of house and predicting next-day average consumption for every half-hour, based on the observed data.
- Find out any patterns of electricity consumption, and whether there are any repetitive ones.
- Build a basic user interface which allows consumers to find out what the expected average electricity consumption for the type of house they live in, is.

1.3 Outline

This chapter presents an overview of the project, its motivations and objectives. Chapter 2 gives a summary of related work in the field, introduces the dataset, defines the technique used for fulfilling the objectives of this paper, and describes the parameters of evaluation used. Chapter 3 unveils the process of preparing the data and the analysis carried out in order to extract features from the data, while chapter 4 features the description of a selected naive baseline, as well as a comparison between the main model's results and the naive model's, based on the established evaluation metrics from chapter 2. Lastly, chapter 5 presents the software application sums up the work and hints at possible continuations of the project in the next academic year.

Chapter 2

Background

2.1 Related Work

Electricity consumption has been shown to be affected by multiple factors. In [9], researchers studied groups of houses situated in the north-east of Scotland, around Aberdeen. They conclude that household size, dwelling type and appliance ownership are useful predictors of electricity consumption and their findings hint at a discernible pattern whereby the time and profile of peak-time consumption is, in part, related to household type. Their analysis has shown that couples with children are more likely to own a tumble dryer and a dishwasher and that their washing machine use is significantly higher than couples who did not have children, which may hint at a higher electricity consumption.

[13] showed that dwelling type is an important factor in determining electricity consumption, supporting their argument with facts about heat loss – eloquent for the current paper is that the heat loss in a bungalow (229 W/°C) is greater than in a flat (182 W/°C). Researchers noticed that households in rural locations have much higher average energy use than those in urban locations. A reason given for that is that flats are more energy-efficient than, for example, terraced housing. Another observation is made with regards to the level of income – the energy use is shown to be related to income levels, with energy demand being highest in prospering suburbs and lowest in areas populated by people with constrained financial possibilities. It was reflected that electricity consumption is more strongly correlated with the number of people per household than gas or fuels consumption since disparities in consumption patterns are less pronounced.

Findings in [7] include that older people in Scotland generally structure their day around their favourite television programmes, their carers or activities they were committed to outside their home. It also investigates their willingness to adapt to newer technology and underlines the benefits in financial terms that they reap when they switch to newer technology than the one they have been used to since childhood. They are a representative part of the population – around 16.8% of the Scottish population was aged above 65 in 2021, according to Scotland's Census. It is, therefore, useful to know their patterns of using electrical appliances and other electric systems in their

homes.

Research in the area has mainly focused on the individual household in an effort to find ways to predict more and more accurately the consumption of one household. There has been a wide breadth of research illustrating state-of-the-art methods, mainly based on deep learning, which hypothetically improve accuracy over more traditional machine learning methods, such as linear regression, or conventional time series models, such as ARIMA. In [33], the authors obtain a mean absolute percentage error (MAPE) of 35% via employing a bi-directional GRU architecture. In [18], the authors use a 2-layer LSTM architecture and obtain MAPE of 44%. In [4], the authors combine LSTM and CNN in a model which has a 40% MAPE, compared to a MAPE of 44% with just the LSTM alone. In [6], the authors optimize a CNN by pre-clustering of customers into similar profiles, thus reducing the number of forecasting models required in a practical DSR implementation, and report a MAPE of 39%. In [30], the authors built an echo state network and achieved MAPE's of 32% to 40%. However, despite all these efforts, the accuracy of forecasting electricity consumption remains around 30-45%, as reported using the MAPE [10]. This is mainly due to random human behaviour, which no predictive model can learn.

The presented review of past research underlines the best methods for predicting accuracy at the household level. Given the breadth of data available, this paper attempts to look more closely at electricity consumption patterns and establish any differences between different categories of households. Due to the focus being distributed between both data analysis and model creation, the model that will be constructed will be a traditional one, as described in section 2.4. It will make use of the findings about correlations in the data in order to create a 48-point forecast covering 24 hours at a resolution of 30 minutes of the aggregated power demand by household type, using past data.

2.2 Description of the Data

This project uses data from the IDEAL (Intelligent Domestic Energy Advice Loop) project [28]. The dataset contains electricity and gas data collected between August 2016 and June 2018, with the help of sensors, which were placed in households for at least 55 days and a maximum of 673 days, with a mean of 286 days and a median of 267 days. Homes were required to have a gas central heating system, meaning that their primary source for heating was gas. Most of the rooms in the homes had radiators. 39 of the homes were part of the enhanced group, meaning some additional sensors were placed there.

There are three main types of data: Metadata, Sensor Data and Survey Data. The metadata contains the relevant socio-economic background of the homes and occupants who participated in the study. Sensor data contains time-series data for electricity and gas, collected by the sensors placed in the homes. Lastly, the survey data contains feedback given by the occupants throughout the time of the installed sensors.

Figure 2.1 depicts statistics of the metadata. One may notice that there is some balance in the selection across all depicted categories, although there are some cases where a

majority for one certain class may be identified. For example, if one looks at the stats for ”era of construction”, it is visible that there is a preponderant number of homes built between 1850 and 1899. These graphs will be considered when forming categories of houses in chapter 3.

Figure 2.2 showed when sensor readings were expected in the time frame, and a ratio of the number of responses received to the number of responses expected. The blacks indicate no responses received, while vertical black lines indicate data receiving issues across the whole platform. The many blacks present in the plot indicate that missing data needs to be addressed before any analysis is done.

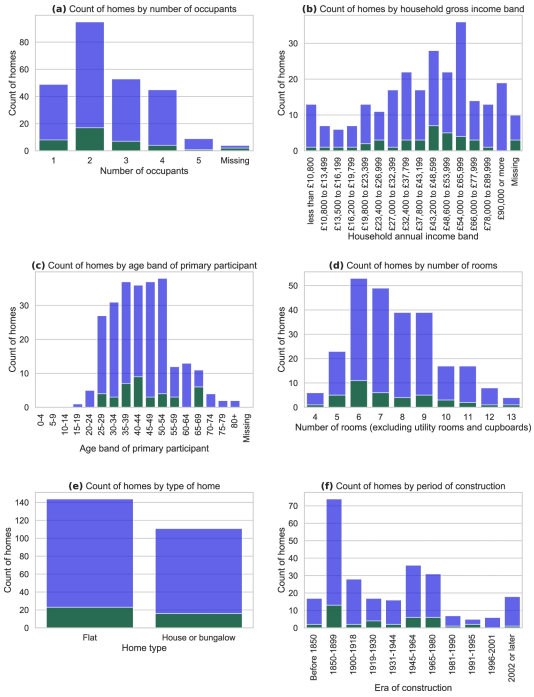

Figure 2.1: Metadata stats. The blue bars represent the total number of homes that fulfill the criteria. The green bars inside are to underline how many of those homes are part of the enhanced group. [28]

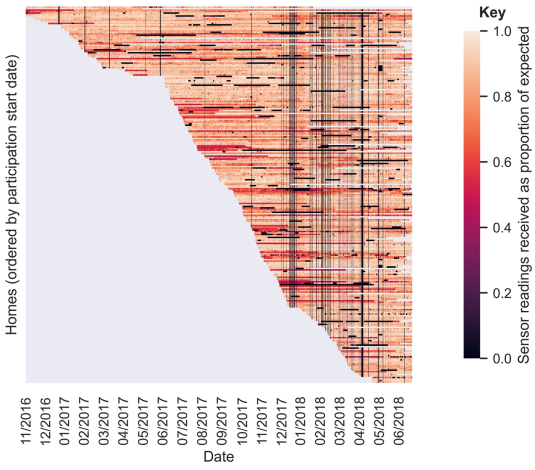

Figure 2.2: Propagation rates of sensor reading. [28]

This project makes use of metadata and electricity sensor data of apparent power. The latter one is collected at a frequency of 1 Hertz using 2 clamps, and is in Watts [28]. Figure 2.3 shows two snapshots of the electricity data collected from one of the homes.

In the picture on the left, one can notice the electricity consumption over all the period during which the home took part in the project, sampled at a 30-minute rate. The blue line that goes up and down the y-axis and along the x-axis shows the consumption across time for a period of just over 1 year and a half. The image also depicts white areas on the graph, meaning there was no data registered from the sensor in that house during those periods.

The picture on the right illustrates electricity data from the same house, on a random

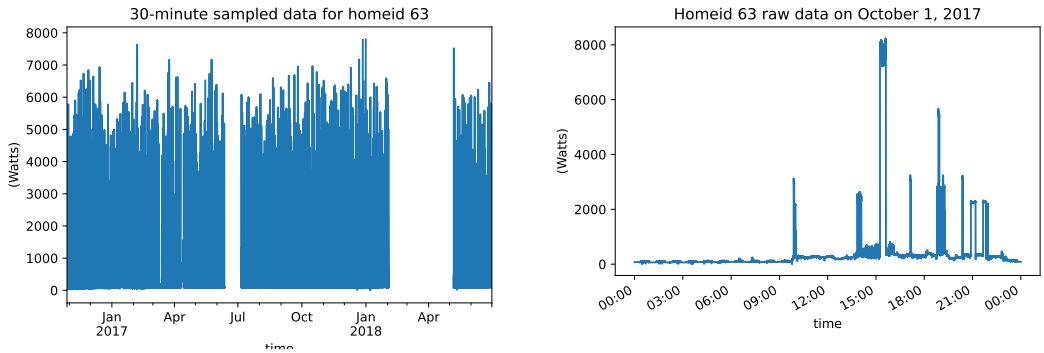

Figure 2.3: Data from one of the homes. [28]

day in the calendar. This is raw data, at the sample rate of 1 second, as given by the sensor, which is why the blue line looks fuzzy and, hence, the data looks rather noisy. One may notice that for large parts of the day, the electricity consumption stays close to 0, especially at night. There are peaks of electricity consumption, which can be noticed in the morning (around 9am to 10am), in the afternoon (around 2pm to 4pm) and in the evening (around 7pm to 11pm).

2.3 Forecast Model – ARIMA

ARIMA is a statistical model and is one of the two most widely used approaches to time series forecasting, alongside exponential smoothing [17]. Also referred to as the Box-Jenkins methodology, ARIMA stands for Auto-Regressive Integrated Moving Average. It aims to describe the autocorrelations in the data and has 3 parameters:

- AR parameter p represents the order of the auto-regressive process. Such process is a regression of the variable against itself, meaning that it predicts the variable of interest using a linear combination of the p most recent values of the variable [17].
- I parameter d represents the order of difference. Differencing a series means calculating the difference between consecutive terms, i.e. $T_k = T_k - T_{k-1}$, and cutting out the first term. Parameter d is how many times the process of differencing needs to be applied on the time series until it becomes stationary. A stationary time series is one whose properties do not depend on the time at which the series is observed. Thus, a time series is not stationary if it contains trends or seasonality. Note that a time series can have a cyclic behaviour but no trend or seasonality, which means it is stationary. This is because the cycles have variable lengths, making it impossible to guess where the peaks and troughs of the cycles will be.
- MA parameter q represents the order of the moving average process. This process uses the q most recent forecast errors in a regression-like model. The errors are not observed, though, which makes this a little different from traditional regression models.

Note that if d is not 0, then differencing is done before calculating any predictions. The future value of a variable is expressed as a linear combination of past values and past errors:

$$Y_t = \phi_0 + \phi_1 * Y_{t-1} + \phi_2 * Y_{t-2} + ... + \phi_p * Y_{t-p} + \varepsilon_t - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - ... - \theta_q * \varepsilon_{t-q}$$

where:

- Y_t is the variable being predicted at moment t;
- ε_k for $t-q \leq k \leq t$ are independent, identically distributed error terms with mean 0;
- ϕ_i and θ_j are the coefficients, $0 \leq i \leq p$ and $1 \leq j \leq q$;
- p and q are integers that are referred to as auto-regressive and moving average parameters.

There are three stages of building an ARIMA model: identification, estimation and validation [26]. In the first stage, one or more ARIMA models are selected. The chosen model is the one for which the autocorrelation (ACF) and the partial autocorrelation (PACF) functions are similar to ACF and PACF calculated based on the sample of observations. In the second stage, the estimates of the parameters of the ARIMA model are obtained, provisionally chosen at the stage of identification. In the third stage, tests are performed to determine if the model estimated is statistically adequate.

Different variants of the model have been developed along the time:

- The Seasonal ARIMA (SARIMA) was developed to help the ARIMA model account for seasonality in data [17]. A seasonal ARIMA model is denoted as SARIMA(p,d,q)(P,D,Q)[m] and includes, in addition to the basic ARIMA model parameters (p, d, q), the parameters m, P, D, Q. m is the length of the seasonal period. For example, if patterns repeat in yearly manner in a dataset which contains monthly data, then m is equal to 12. P, D and Q are equivalent to the non-seasonal components of the model, except they involve backshifts of the seasonal period. One could say that ARIMA is a SARIMA with m=1.
- ARIMAX is an ARIMA with exogenous variables. An exogenous variable is defined as a variable which appears in the model, but is not explained by the model, i.e. it is given to the model. [22]
- SARIMAX is a seasonal ARIMA model with exogenous variables.

The experiments will make use of the ARIMAX and SARIMAX models.

2.4 Evaluation Metrics

The evaluation metrics used in this paper are standard and used in many papers on similar topics.

Root Mean Squared Error

The root mean squared error (RMSE) is a scale-dependent metric [17], always non-negative and for which smaller values indicate better performance. Long right tails in the data show in the results of RMSE, since large individual errors have a big effect on the RMSE. Minimising the RMSE leads to forecasts of the mean.

The metric will be used to calculate an average error for the day, as well as an average of the errors for all test days. Since there are 48 half-hours in a day, the RMSE for a single test day is:

$$RMSE_d = \sqrt{\frac{1}{48} * \sum_{i=1}^{48} (\hat{y}_{d,i} - y_{d,i})^2}$$

where i is the interval in the forecast horizon, d is a day in the test period, $\hat{y}$ is the interval forecasted power, y is the actual interval power. The RMSE for the entire test period is then:

$$RMSE = \frac{1}{D} * \sum_{d=1}^{D} RMSE_d$$

where D is the total number of days in the test period.

Mean Absolute Error

The mean absolute error (MAE) is another scale-dependent metric [17], which differs from RMSE in that it does not penalise larger errors as excessively, but, instead, reports the average absolute error. A forecast method that minimises the MAE will lead to forecasts of the median.

The metric will be used in similar fashion to the RMSE. The MAE for a day is:

$$MAE_d = \frac{1}{48} * \sum_{i=1}^{48} |\hat{y}_{d,i} - y_{d,i}|$$

And the MAE for the entire test period, consisting of D days, is:

$$MAE = \frac{1}{D} * \sum_{d=1}^{D} MAE_d$$

Mean Absolute Percentage Error

The mean absolute percentage error (MAPE) is a unit-free metric, frequently used to compare forecast performance between data sets [17]. One behavioural issue that might be encountered is with consumption averages close to 0, values to which this metric does not respond well.

MAPE will be used in similar fashion to the other 2 metrics. The MAPE for a day is:

$$MAPE_d = \frac{100}{48} * \sum_{i=1}^{48} \frac{|\hat{y}_{d,i} - y_{d,i}|}{y_{d,i}}$$

And the MAPE for the whole test set containing D days is:

$$MAPE = \frac{1}{D} * \sum_{d=1}^{D} MAPE_d$$

Chapter 3

Data Preparation

3.1 Data Pre-processing

The aim of pre-processing data was to split homes into categories and produce train and test sets for each type of home, containing the relevant readings and information, with all incomplete, corrupted, incorrect, inaccurate and irrelevant data removed. Each row in the final data frames represent a 30-minute average value for a certain timestamp. The data that is used to produce this row consists of readings from 5 to 10 homes of the respective type, which are downsampled to 30-minute intervals. In summary, for each home, an average value is calculated for every 30-minute interval, and that value is then used in the calculation of the value in the final data frame for the specific type of home.

3.1.1 Categorisation and Aggregate Processing

The homes are split in categories based on 3 factors: type, number of people living in, and the build era. The factors were selected taking into account the research shown in related work, while the splitting limits were established with the statistics shown in Figure 2.1 in mind. The scope was to get significant categories, but also to have a significant number of homes in each category. Table 3.1 illustrates the categories and the number of homes in each of them.

For each category, a new dataset is created, containing the data sampled at half-hour rate and averaged across 5 to 10 homes for each half hour. The consumption has proved to vary a lot for every category. A sample of the variation is shown in Figure 3.1. It might be explained by different routines in each home, as well as the rather small granularity (30 minutes) at which the data is downsampled. Figure 2.3 is particularly relevant in the explanation of this, as it can be noticed that spikes in electricity usage during a day happen at certain times for a home. It suffices that another home has the same spikes but at different times, for it to show up on this graph with a big deviation from the average consumption in that 30-minute interval. This is also in line with what [5] defines households as, namely heterogeneous customers, who consume different amounts of electricity for different purposes and at different hours of the day.

Number of people	Flat		House	
	Before 1900	After 1900	Before 1965	After 1965
2 or less people	50 homes	48 homes	24 homes	22 homes
3 or more people	27 homes	16 homes	41 homes	23* homes

Table 3.1: Number of homes in each category. There are 4 homes with missing data on occupancy, so they could not be classified in any category.
Note* home with homeid 223 has no issued electric-mains data file, which is the type of data used in this project, so in practice there are only 22 homes analyzed for that category.

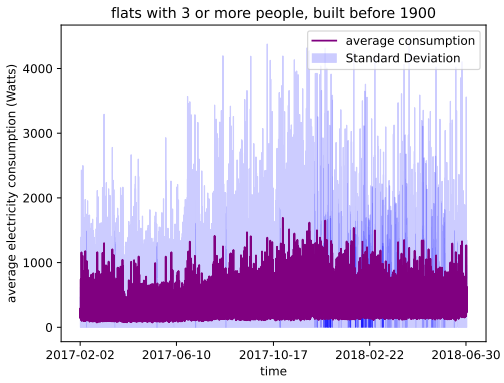

Figure 3.1: Averaged data for one category, together with standard deviation at each timestamp.

3.1.2 Data Cleaning and Splitting

There were two main steps in cleaning the data, each with their substeps. Firstly, the data for each particular house was addressed. In this round, It was firstly checked whether all houses have electricity data ready for analysis. Home with homeid 223 was the single one to miss electric mains readings completely. Once this was done, the next issue that got addressed was missing data. In the original data, for all houses, gaps of 5 minutes or less were interpolated linearly. All data was then downsampled to 30-minute granularity, and any such interval which had less than 5 minutes worth of data to average from, was dismissed.

The data was then grouped in small sets, each set representing a category. For each set, at every timestamp, a maximum of 10 non-null data points were extracted. If there were at least 5 non-null data points, then the values were averaged and the new value would become the representative of the consumption for that timestamp in the new data frame. If there were less than 5 non-null data points, then a null was passed in for that timestamp.

The second round of cleaning followed after obtaining the new data sets, one for each category. Firstly, outliers were considered, although lightly, in that only the very rare

Number of people	Flat		House	
	Before 1900	After 1900	Before 1965	After 1965
2 or less people	516 days	537 days	521 days	333 days
3 or more people	514 days	298 days	368 days	437 days

Table 3.2: Number of days of data available for each category.

high values were deleted. High values that occurred relatively often were kept, due to a willingness that the model would account for such extreme, yet seemingly regular events too. As such, for 3 out of 4 flat types, any average consumption of over 1300 Watts was dismissed. This was also done for houses with 2 or less people living in, built after 1965. For the remaining 3 categories of houses, outliers were considered values that went over 1800 Watts – such values were dismissed. The final step was to fill in the missing data: for each month in the respective year, an average value was calculated, and any null data point in that month would be filled with it.

Each data frame was then split in 2 sets, one for training and one for testing. Data before April 1st, 2018 was used for training, while the data on April 1st and until the end of the period (June 30th), was saved for testing. In ratio terms, this would translate roughly into an 80-20 split.

3.2 Exploratory Data Analysis

Table 3.2 shows how much data each category had left after the data was cleaned.

For the purposes of this paper, experimentation was done only on 3 of the categories which had more than one year's worth of data. That is due to the limited amount of space, as well as the willingness to present meaningful analysis during each step. This section presents the analysis on 3 flat types, hence the exploratory analysis is shown only on the data from the following categories:

- **Category 1:** Flats, with 2 or less people living in, built before 1900
- **Category 2:** Flats, with 2 or less people living in, built after 1900
- **Category 3:** Flats, with 3 or more people living in, built before 1900

One thing to note is that seasons are considered otherwise than normally in Scotland: Spring (March 1^{st} - May 31^{st}), Summer (June 1^{st} - August 31^{st}), Autumn (September 1^{st} - November 30^{th}), Winter (December 1^{st} - February 28^{th}/29^{th}).

Figure 3.2 illustrates, for each category, a blue line representing the average electricity consumption at every half hour, compared to the moving average over the last 7 days, represented by the orange line. One could notice that the moving average is much more stable overall, and hints at a possible electricity consumption average weekly value. However, towards the end of the period, a decreasing trend in usage can be noticed for

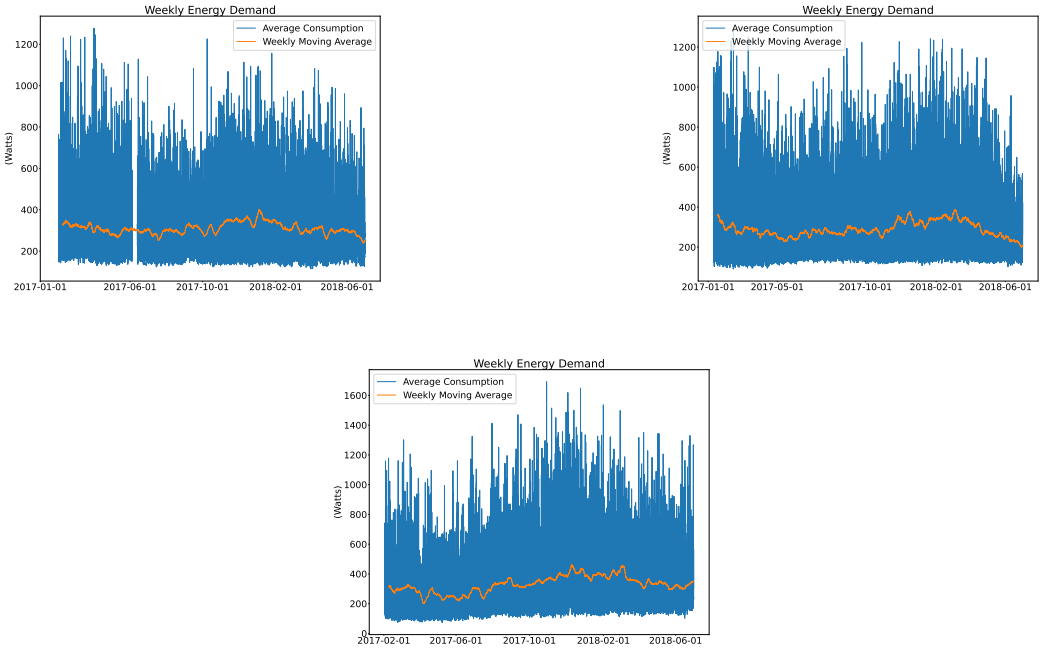

Figure 3.2: Weekly demand compared to consumption per half-hour, for Categories 1, 2 and 3.

Category	Skewness	Kurtosis
1	1.47	6.18
2	1.54	5.85
3	1.49	6.25

Table 3.3: Skewness and kurtosis values for each of the 3 categories.

2 of the categories, which raises the question about what caused the decrease. Notice that the orange line reflects the blue line's direction.

To gather information about the distribution of the data and its symmetry, two values were calculated – skewness and kurtosis. Skewness, in statistics, is the degree of asymmetry observed in a probability distribution [11]. If skewness is 0, then the data is symmetric. On the other hand, kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers [27].

In Table 3.3, one may notice the values for all 3 categories are similar. The skewness values are greater than 0, which means the data is not symmetric and has a right tail – the value of the mean is greater than the median of the set. The kurtosis values are greater than 3, which means the data sets for all 3 categories have a leptokurtic distribution. Such distribution is said to have a wider or flatter shape with fatter tails, resulting in a greater chance of extreme positive or negative events.

A further look at the coefficient of variation reveals more information about the data.

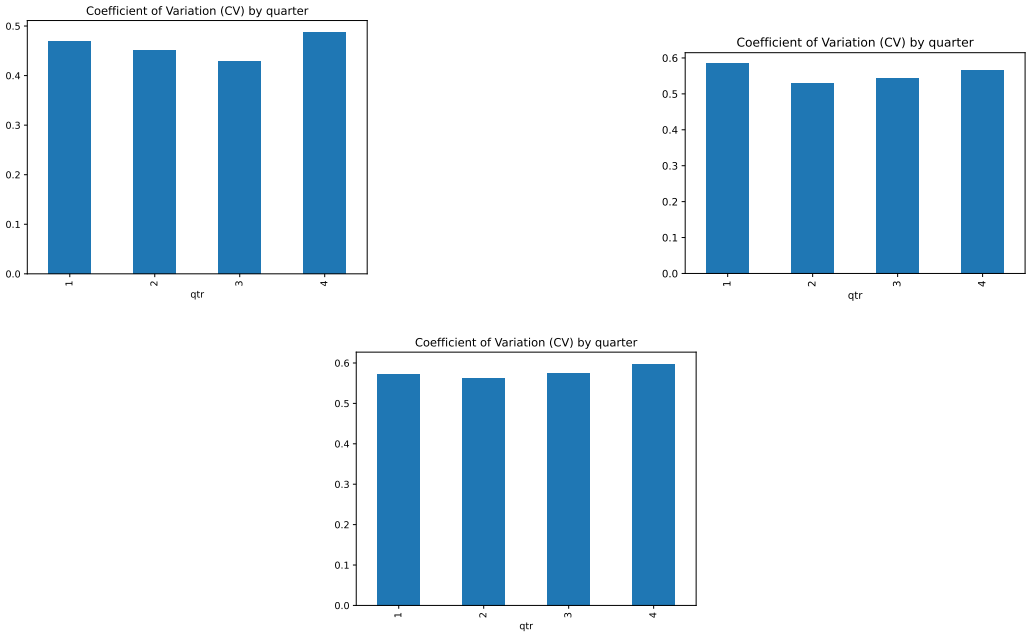

Figure 3.3: Coefficient of variation per quarter of a year for Categories 1, 2 and 3.

Shown in Figure 3.3 is the coefficient of variation by quarter of a year for each category. One could notice that for two of the categories, the coefficient is higher during the last quarter than in the other quarters. A higher coefficient of variation means a higher relative dispersion of data points in the series around the mean. [15]

Analysis on volatility was done too. A look at Figure 3.4 shows how volatile sectors of the data are for each category. The discussion is made around percentiles. The k^{th} percentile, in this case, shows the maximum average consumption for k% of the homes. Volatility is a good indicator of change in consumption patterns, because it allows one to observe how a sector of the market behaves. For example, if the 90^{th} percentile falls, that means that 90% of the houses have a decreased consumption compared to previous period. 10^{th} percentile illustrates the behaviour of the smallest consumers on the market, while the 50^{th} percentile is basically equivalent to the median electricity consumption.

For an ARIMA model, it was explained in section 2.3 that trend and seasonality are two important characteristics which need identified and checked. Figure 3.5 depicts boxplots for each season of the year. Particularly noticeable is that the median, which is depicted through thick horizontal lines running somewhere in the middle of each boxplot, is generally lower during summer and spring. This makes sense, since the period between March 1^{st} and August 31^{st} in Scotland is warm and the sun shines brightly, which eliminate the need for keeping light on inside the house, or for any electric heater to be plugged in. Also noticeable is that the data is skewed and there are many outliers, depicted with black diamonds above the boxplots.

In Figure 3.6, an additive model was assumed in order to check for trend and seasonality. One can observe that no seasonality was found, and the trend is rather cyclical, with

Figure 3.4: Volatility analysis for Categories 1, 2 and 3.

highs and lows happening at irregular times. The residuals show a constant variability for most times, with rare, irregular peaks happening here and there.

Heteroskedasticity analysis is concerned with residuals. The presence of heteroskedasticity implies the absence of homoskedasticity – in other words, the variance in electricity consumption is not constant. This is shown in Figure 3.7, where the variance per month is depicted against the variance per week for the 3 categories. It can be observed that there is more variance from one week to another than from one month to another. However, the errors in neither case are constant, for any of the categories.

Figure 3.8 draws attention over the distribution of the data. For each category, a probability plot was drawn, to check whether the electricity consumption values follow a normal distribution. It is visibly not the case for any of the 3 categories.

A density plot is concerned with the distribution of data. It uses a kernel to approximate the probability density function of the average electricity consumption. Such plots can be seen for each category in Figure 3.9, where the darker blue line represents the probability density function. Additionally, in each plot depicted is the mean value for average consumption with a red horizontal line, as well as what would be the limits of a 95% confidence interval for a normal distribution – that is, the values that are 2 standard deviations away from the mean. It is interesting to note that most values actually fall in this interval, for all 3 categories.

3.3 Stationarity Evaluation

Table 3.4 depicts the results of an Augmented Dickey-Fuller test carried out on the data sets for each of the 3 categories. The test evaluates whether a time series has a

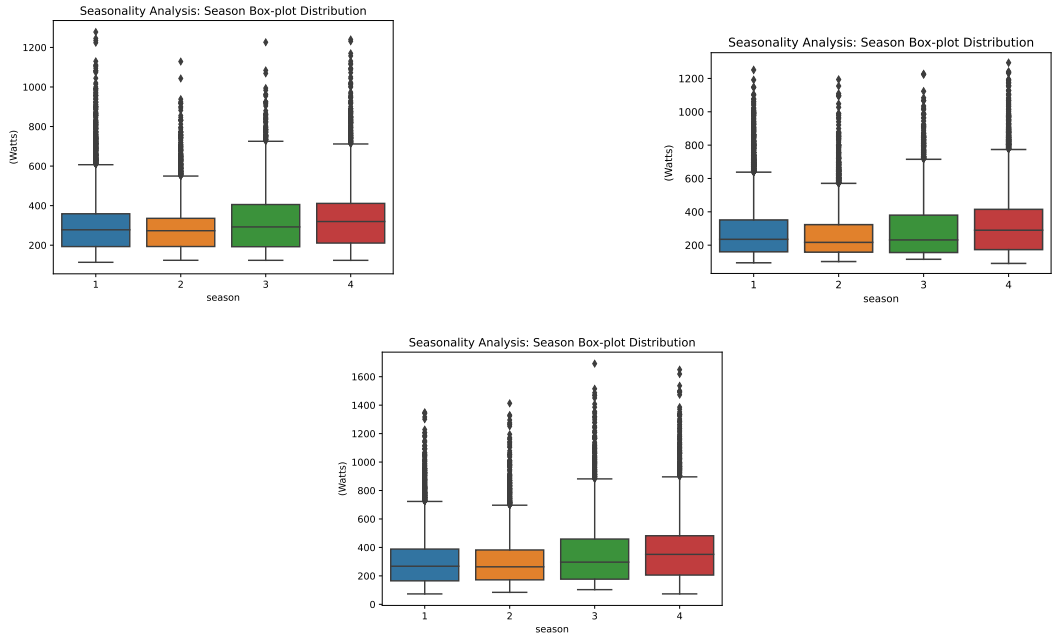

Figure 3.5: Distribution of data on seasons for Categories 1, 2 and 3. Season 1 is spring, season 2 is summer, season 3 is autumn and season 4 is winter. Seasons are considered in a different manner than normally done in Scotland.

Category	ADF Stat	p-value	CV 1%	CV 5%	CV 10%	result
1	-11.796	0.000000	-3.431	-2.862	-2.567	stationary
2	-11.247	0.000000	-3.431	-2.862	-2.567	stationary
3	-7.98	0.000000	-3.431	-2.862	-2.567	stationary

Table 3.4: Augmented Dickey-Fuller test results for each of the 3 categories. CV stands for ”Critical Value”.

unit root – meaning it is non-stationary. If this hypothesis is rejected, then the data is stationary. The hypothesis can be rejected if the ADF statistical number is smaller than all critical values, and the p-value is smaller than a set threshold – in our case, 0.05. The analysis done in previous sections have already hinted at the answer, and the results of the test are just confirming the expectations – the data is stationary for every category, significant to the critical value of 1%.

Another way of testing stationarity is by checking the autocorrelation function (ACF) and the partial autocorrelation function (PACF) plots. Any regular correlation may indicate seasonality. The ACF plot shows the autocorrelations, which measure the relationship between y_t and y_{t-k} for different values of k [17]. If y_t and y_{t-1} are correlated, then $yt-1$ and y_{t-2} must also be correlated. However, there might exist correlation between y_t and y_{t-2} just because they are both connected to y_{t-1}.

To find out if that is the case, one can use the PACF plot. It measures the relationship between y_t and y_{t-k} after removing the effects of lags 1,2,3,...,k-1. Each partial au-

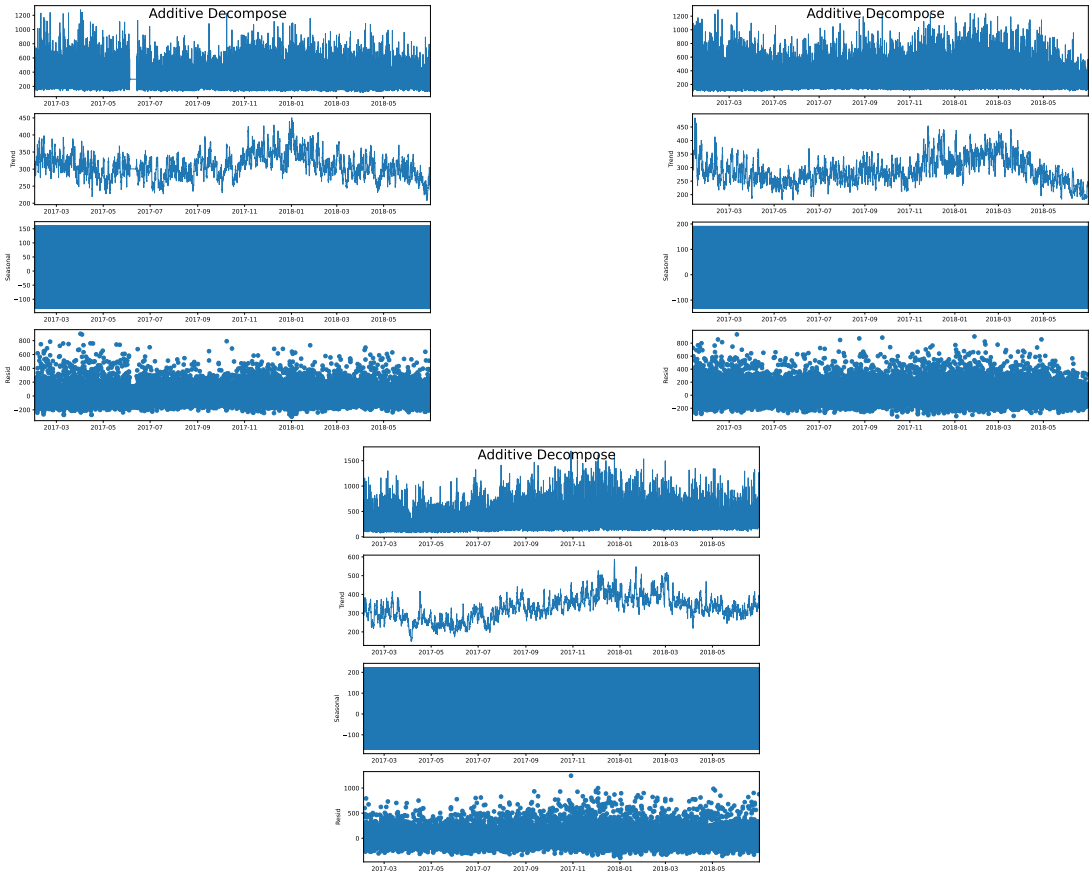

Figure 3.6: Additive decompose for Categories 1, 2 and 3.

tocorrelation can be estimated as the last coefficient in an autoregressive model, a fact which will be considered when inputting model parameters in chapter 4. Specifically, α_k, the k^{th} partial autocorrelation coefficient, is equal to the estimate of ϕ_k in an AR(k) model.

Figures 3.10 and 3.11 show the ACF and PACF plots for all 3 categories. The ACF plots unveil positive correlations between the time series and the first 8 lagged values, and negative correlations for the rest of the lagged values up to the 48^{th} one, which has again positive correlation. There is a certain seasonality that can be noticed into the data, for all categories.

The PACF plots show that the first two lagged values have significant positive correlation with the current value, and the rest of the lagged values having mostly negative, significant correlation to the value at the current moment. There is again a noticeable seasonality, with correlation becoming significantly positive around lag 48.

The observed patterns call for an evaluation of the differenced data. And, indeed, the autocorrelation plots for the differenced series looks much better. In Figure 3.12, one can observe the autocorrelations are positively significant only for the first few lagged values. There is a spike around the 48^{th} lag again, but that is about the only other significant correlation.

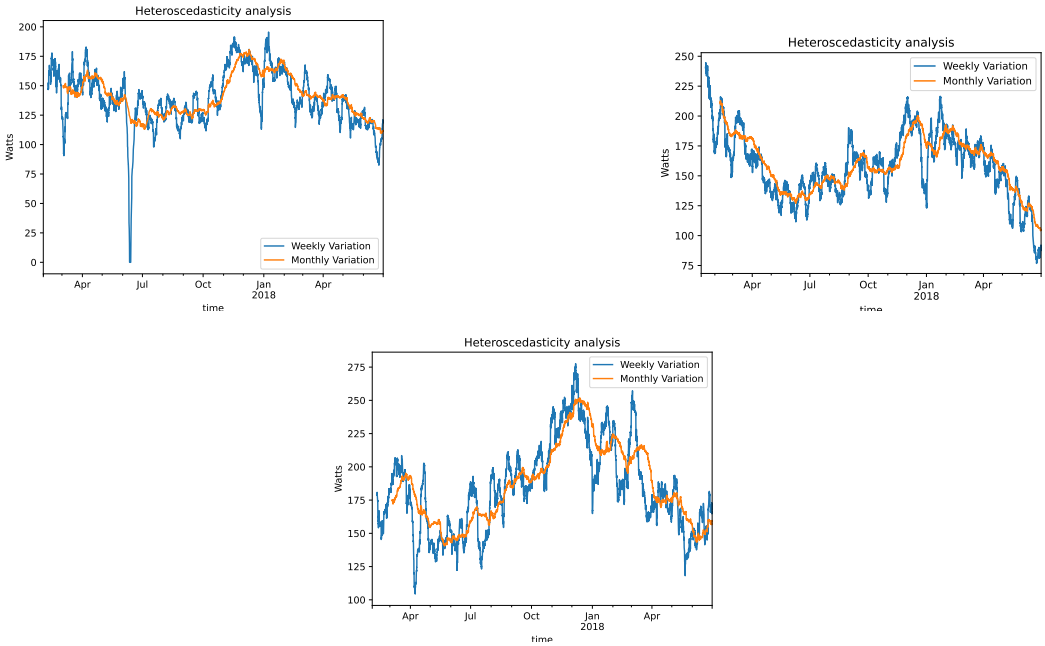

Figure 3.7: Heteroskedasticity analysis for Categories 1, 2 and 3.

Figure 3.13 shows that, indeed, differencing reduced correlations between the current lag and the previous ones, which is indicative that differencing was the right thing and that in fact, the initial series might be non-stationary. There are significant autocorrelations at lags situated at indices which are multiple of 48, and a significantly positive correlation with lag 49 and 97.

The results above indicate that constructing a seasonal model is actually worth considering, so the next section will compare a seasonal model against a non-seasonal one.

3.4 Feature Selection

Table 3.5 shows a dictionary of abbreviations used and the full name of the corresponding feature. Table 3.6 presents how the average electricity consumption attribute ”avg” relates to the other features in each data frame, calculated using the Pearson Correlation Coefficient. It can be seen that, generally, the hour at which the prediction is done has a high correlation with the average consumption. There is some consistency across all categories in that the electricity consumption for the past 3 days, or week, or month, correlate with the result quite well and in similar terms. The same can be said about the season during which the measurement is taken.

Different types of feature engineering and preprocessing were carried out. Firstly, temporal data was extracted from the time index, such as the hour, the day as a number from 0 (Monday) to 6 (Sunday), the year’s quarter, the season (Spring, Summer, Autumn, Winter – as described at the start of section 3.2). Some other features were created, such as a boolean whether the timestamp is situated during what is considered

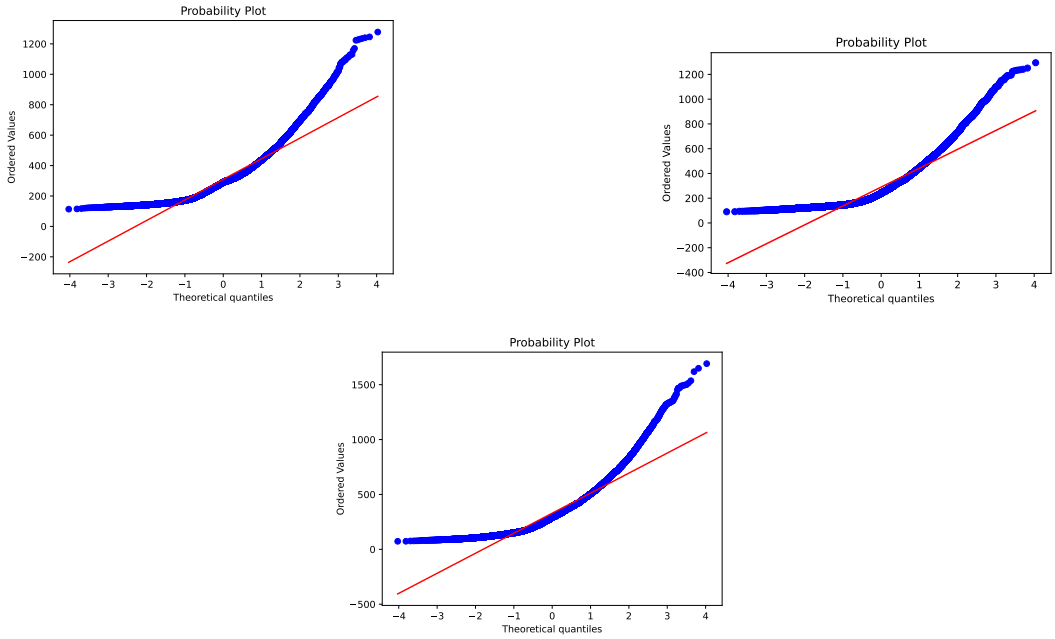

Figure 3.8: Probability plot for Categories 1, 2 and 3. Blue points represent actual average consumption values, while the red line illustrates a normal distribution.

business hours in the UK, a boolean specifying whether the timestamp occurs while the sun is up and is daylight, and an integer to establish whether the timestamp is situated during the working week (Monday to Friday, marked with 0) or in the weekend (1 for Saturday and 2 for Sunday).

The features the resulted from exploring the data. like the quantiles, the moving averages, the moving standard deviations were excluded in the final model. The argument is that ARIMA is supposed to calculate moving averages, so there is no need to feed in that kind of information.

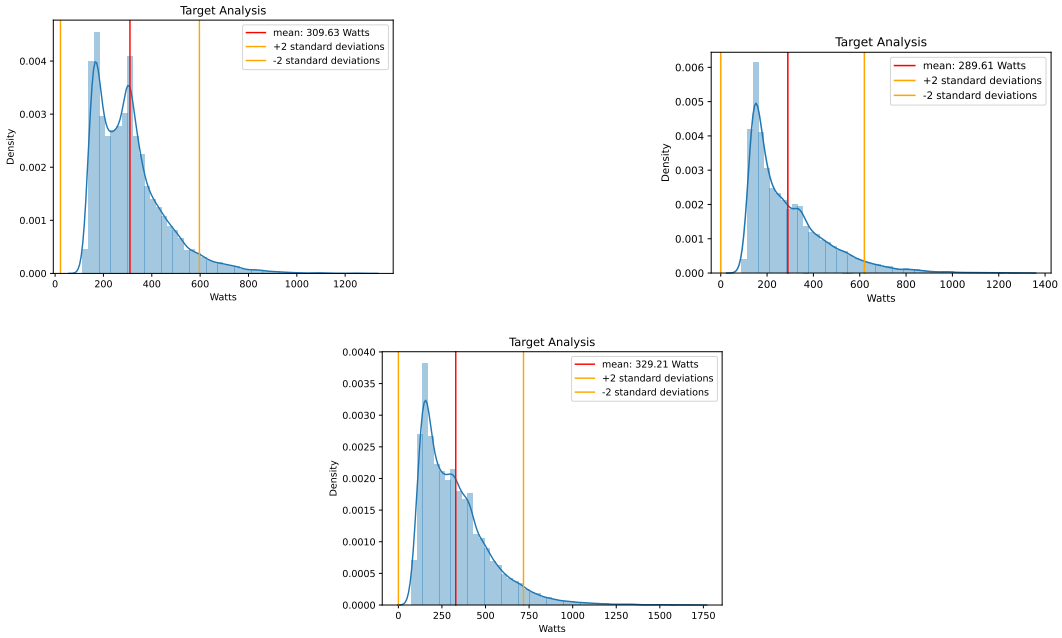

Figure 3.9: Target analysis for Categories 1, 2 and 3.

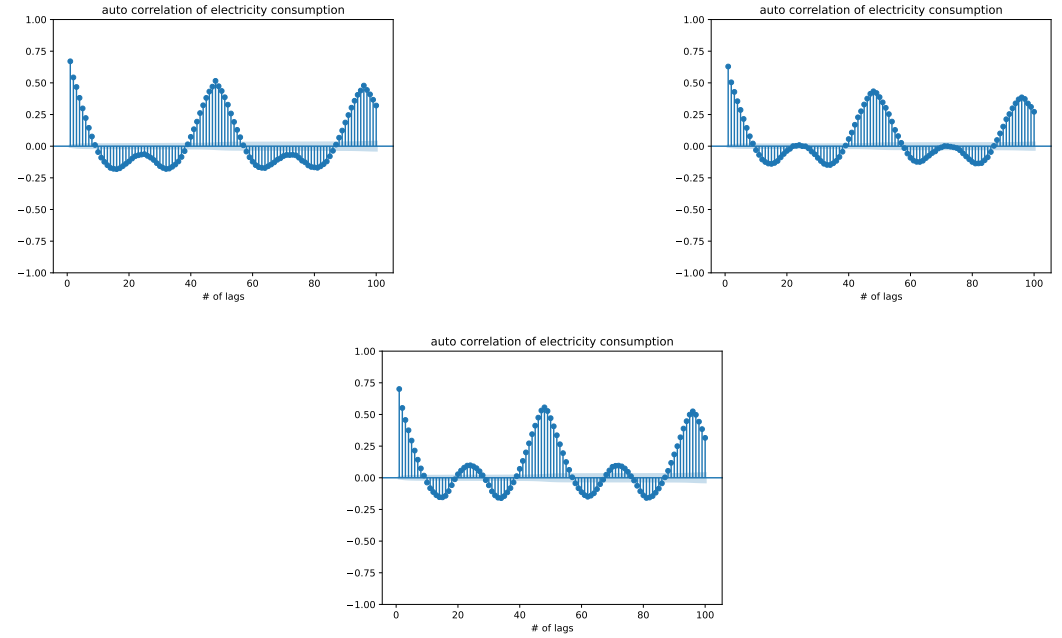

Figure 3.10: Autocorrelation Function for Categories 1, 2 and 3.

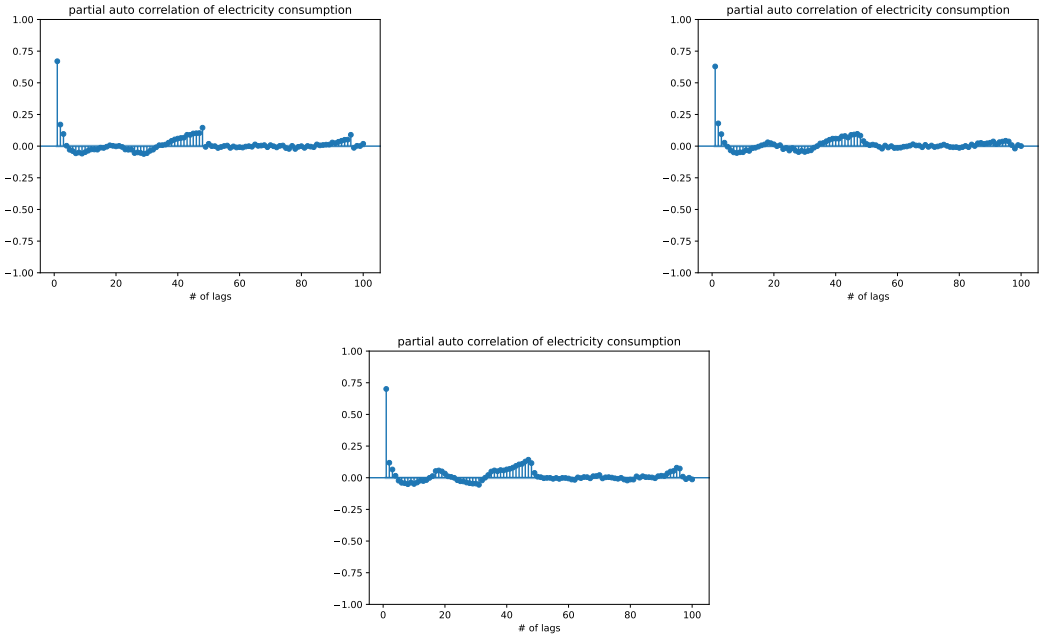

Figure 3.11: Partial Autocorrelation Function for Categories 1, 2 and 3.

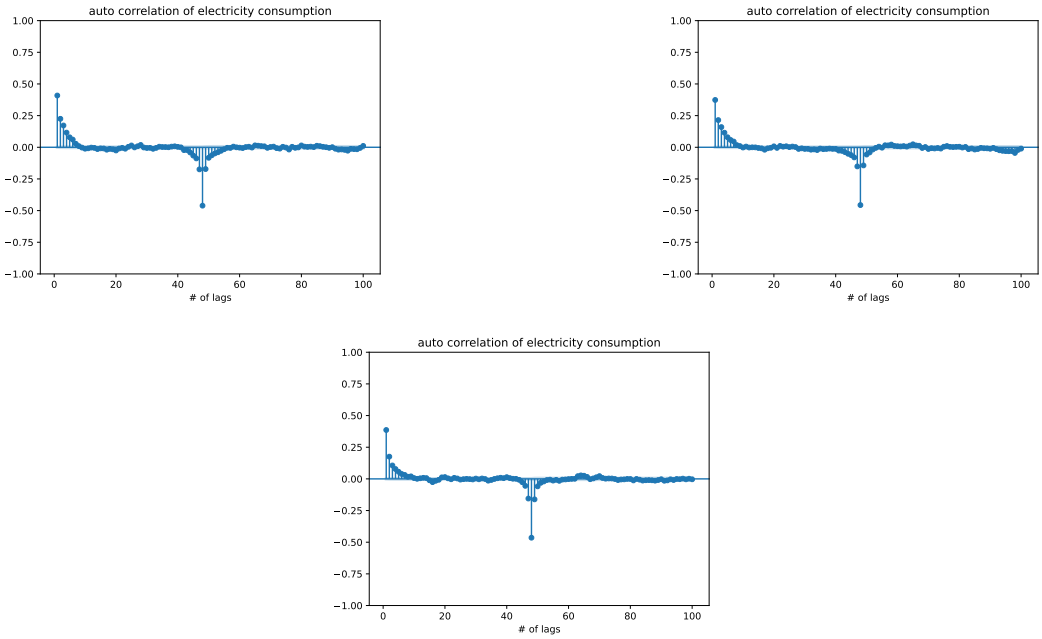

Figure 3.12: Autocorrelation Function for the differenced series of Categories 1, 2 and 3.

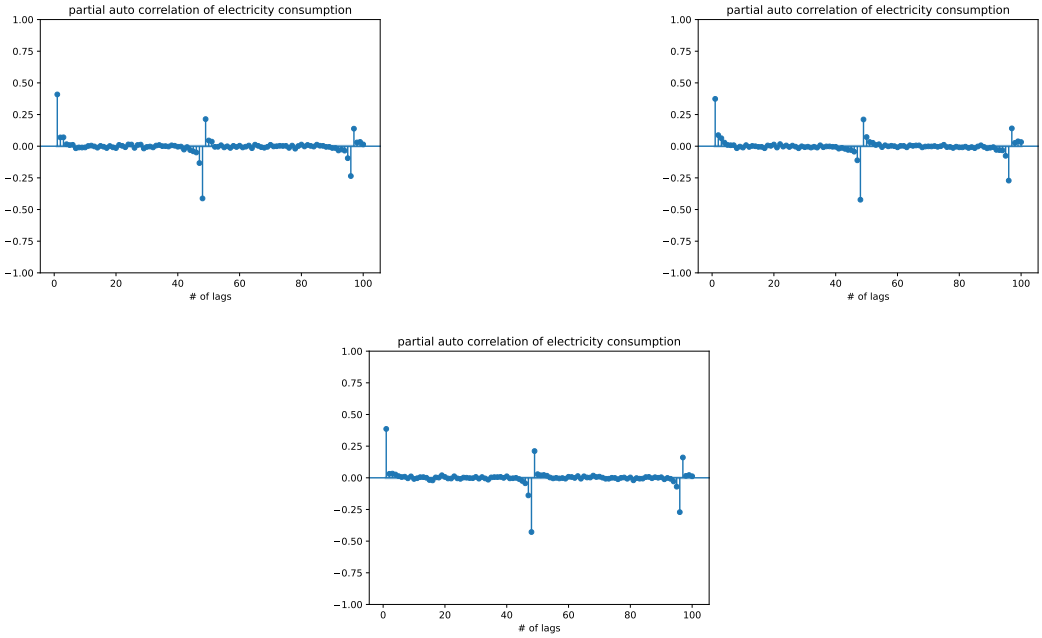

Figure 3.13: Partial Autocorrelation Function for the differenced series of Categories 1, 2 and 3.

Feature name	Feature abbreviation
Electricity consumption average for 30 minutes	avg
Hour in which the half-hour happens	hour
Boolean indicating whether it is daylight outside	daylight
Moving average for the past 3 days	movave_3
Moving average for the past 7 days	movave_7
Moving standard deviation for the past 3 days	movstd_3
Moving standard deviation for the past 7 days	movstd_7
Moving standard deviation for the past 30 days	movstd_30
Moving average for the past 30 days	movave_30
90^{th} percentile	q90
Boolean indicating whether it is a business hour	business hour
season of the year	season
50^{th} percentile	q50
day label	day
Weekend indicator	weekend
Quarter of the year	qtr
Month of the year	month
10^{th} percentile	q10

Table 3.5: Pearson Correlation for the target attribute ”avg”.

Feature abbreviation	Category 1	Category 2	Category 3
avg	1.000000	1.000000	1.000000
hour	0.398291	0.449920	0.392653
daylight	0.282351	0.128674	0.216693
movave_3	0.184867	0.223332	0.282507
movave_7	0.169504	0.212145	0.271215
movstd_3	0.122278	0.173668	0.227754
movstd_7	0.126542	0.165787	0.229864
movstd_30	0.129876	0.169095	0.226064
movave_30	0.141734	0.200418	0.252342
q90	0.148204	0.187904	0.253318
business hour	0.132808	-0.013874	0.063711
season	0.105279	0.111264	0.157557
q50	0.149898	0.189334	0.255239
day	0.038494	0.049351	0.057690
weekend	0.035717	0.074797	0.073459
qtr	0.012968	-0.035171	0.049134
month	0.013200	-0.036416	0.053087
q10	0.004923	0.120443	0.209923

Table 3.6: Pearson Correlation for the target attribute ”avg”.

Chapter 4

Results

4.1 Model Parameters

4.1.1 ARIMAX

The first try was to fit non-seasonal models on the data for each category, due to the results given by the Augmented Dickey-Fuller test. For this, the python library pmdarima was used, due to having an auto-arima function, which performed a step-wise search to find the best fitting parameters for the model. In Table 4.1, the parameters for each model can be seen, as well as coefficients to each feature and some scores, like Akaike Information Criterion (AIC), which was the main criteria for choosing the best model – whichever model had the smallest AIC score was considered the best. AIC is an estimator of the model's prediction error, relative to each of the other models [3].

Notice that the parameters are similar from one category to another. What differs is how well the model fits the data for each category – in this case, category 1 has the smallest AIC score among the 3 categories. This theoretically means that the prediction error will be the smallest for category 1.

4.1.2 SARIMAX

The pmdarima library did not work at identifying seasonal models in the series. That might be due to the library not being well optimized for long seasonality – as observed in the auto-correlation plot in Figure 3.10, the seasonality period is 48 points long. Hence, there was need for a more canonical approach – the statsmodels library. This one did not have a function like auto-arima, so 8 experiments were carried out by hand for each category. The model parameters for testing were chosen with respect to the observations made on the autocorrelations and partial autocorrelation plots in section 3.2. Seasonal period was set to length m=12, the order of seasonal differencing was always set to 1, and so was the seasonal moving average always set to 1. The rest of the parameter settings depended from case to case. The models with the lowest AIC score for each category are summarised in Table 4.2.

Category	Model	AIC Score
1	ARIMAX(3,0,3)	247415.3
2	ARIMAX(3,0,3)	267925.17
3	ARIMAX(3,0,2)	256986.52

Table 4.1: Model parameters and AIC scores for non-seasonal ARIMAX.

Category	Model	AIC Score
1	SARIMAX(1,0,2)(1,1,1)[48]	243024.71
2	SARIMAX(1,0,2)(1,1,1)[48]	264473.4
3	SARIMAX(1,0,2)(1,1,1)[48]	252156.38

Table 4.2: Model parameters and AIC scores for seasonal ARIMAX.

4.2 Performance Analysis

4.2.1 ARIMAX results

Firstly, the model's ability to fit the data is evaluated, i.e., capture the information in it. A residual is known as the difference between the fitted value and the actual value. There are 3 properties of the residuals which are checked:

- The residuals have 0 mean. If that does not happen, then it means that predictions of the model are biased.
- The residuals are uncorrelated. If there is any significant correlation, that means there is information left in the residuals, which has not been captured by the model.
- The residuals are normally distributed and/or have constant variance. This would help at calculating the prediction intervals better.

Figure 4.1 helps assess these properties. It contains three groups of plots, one for each category. Each group contains four plots, which are being explained below.

The first plot (top-left) shows the residuals at each timestamp, on a standardised scale. It is useful for identifying any seasonality or trend in the residuals, as well as eye-balling the mean of the data. Ideally, neither trend nor seasonality are observable, and so is the case here. The residuals also seem to float around 0.

The second plot (top-right) shows the distribution of the residuals against a normal distribution, as well as a histogram of the predicted values, again on a standardised scale. Coloured with orange is the density of the residuals, which is compared to a normal distribution, represented by a green line. A normal distribution is a distribution with mean 0 and standard deviation 1. This plot also helps at clarifying whether the mean is 0 or not, which in this case seems like it is slightly off. There is, hence, a slight bias in the model. It is also visible that residuals do not follow a normal distribution, which is further clarified by the 3rd plot (bottom-left). That shows how skewed the

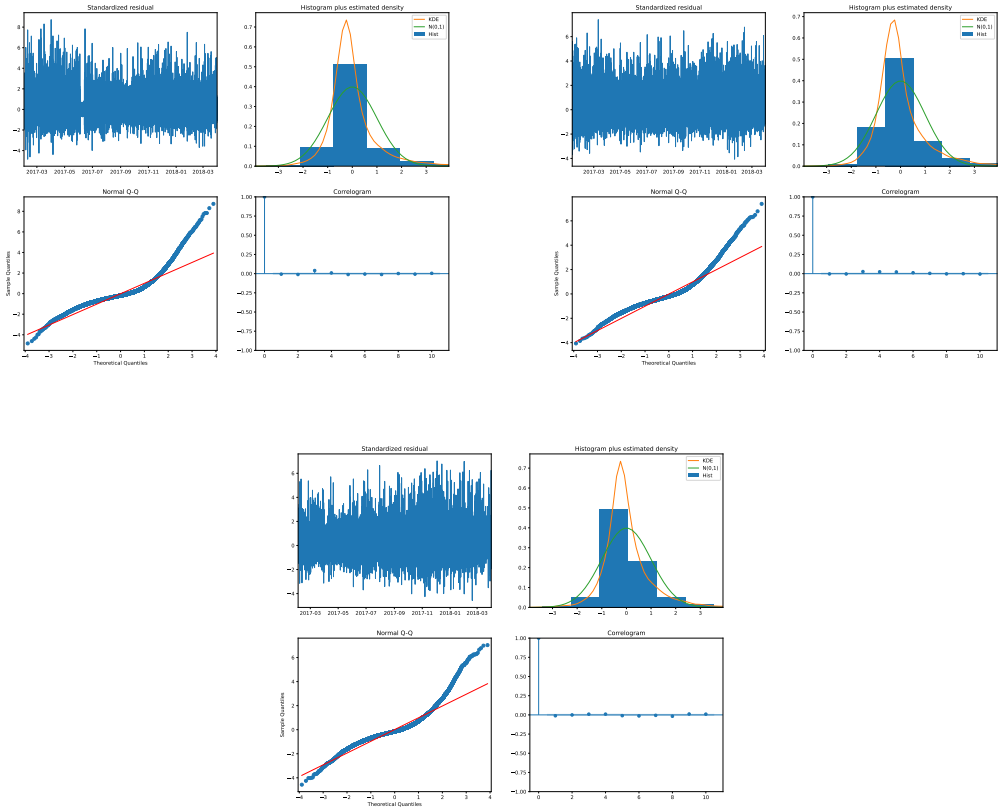

Figure 4.1: Diagnostics of predicted values by ARIMAX for each of the 3 categories, in order: category 1, category 2, category 3. Analysis of each diagnostic is presented in section 4.2.

predicted values are against a normal distribution. It can be observed that values in the larger spectrum are skewed, which means the model's predictions are heavily right-tailed.

Lastly, the 4th plot (bottom-right) shows how autocorrelated the residuals are. There is no significant autocorrelation, which means that the model captured every bit of information it could.

Table 4.3 shows that the series of residuals is stationary according to the Augmented Dickey-Fuller Test, which is a good sign that the errors do not correlate with one another in any way.

Secondly, the actual predictions are evaluated. Figure 4.2 shows the model's predictions compared to actual values on the testing set. How the split was done is described in chapter 3. One may notice the model captures the general shape, however it fails to accomodate for the extreme events. The density of the data, presented in Figure 3.9, might help at explaining why the predictions mostly fall between 200 and 400 Watts – most values are situated in that region.

Another observation is that the gaps in data are not captured. This is expected, since

Category	ADF Stat	p-value	CV 1%	CV 5%	CV 10%	result
1	-22.113	0.000000	-3.432	-2.862	-2.567	stationary
2	-7.082	0.000000	-3.432	-2.862	-2.567	stationary
3	-8.5	0.000000	-3.432	-2.862	-2.567	stationary

Table 4.3: Augmented Dickey-Fuller test results for the residuals of the ARIMAX model, when applied on each of the 3 categories. CV stands for ”Critical Value”.

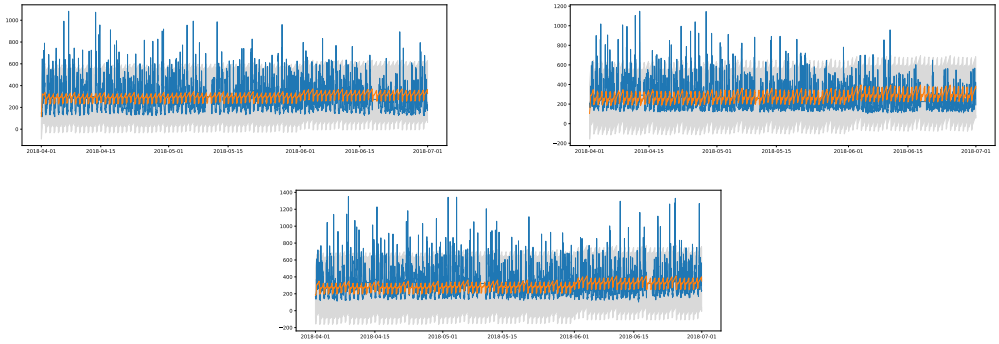

Figure 4.2: Results plot for each of the 3 categories, in order: category 1, category 2, category 3. Blue line represents actual consumption, orange line is the ARIMAX model's prediction, the grey hue is the 95% confidence interval.

they do not occur often or at regular times. The prediction line also fails to follow the blue line towards the end of the testing period – that might have to do with the model's inability to predict on long periods of time, which is a known issue of ARIMA. Despite it representing only 20% of the whole set, there are still 3 full months (April, May, June) which the model has to predict for.

Table 4.4 sheds more light on the matter. It illustrates the correlation between the forecasted values and the other attributes. The predictions here are represented as attribute ”forecast”. One can see that, compared to the actual values, represented by attribute ”avg”, the correlation is only about 40%, which is quite low. The highest correlations are with the ”hour” and ”daylight” attributes, which seem to have been the most valuable features out of those fed to the model. It is interesting to note that correlation is negative with all moving averages and standard deviations, as well as percentiles, which is contrary to the positive correlations that the attribute for actual values has, as shown in section 3.4. This demonstrates the model's inability to capture temporal dependencies on relatively long periods of time.

4.2.2 SARIMAX results

Firstly, the model's ability to fit the data is evaluated. Just like for the ARIMAX model, Figure 4.3, containing 3 groups of images, each of 4 plots, helps asses the properties of the model's residuals. It can be seen that, just like with the non-seasonal model, the residuals seem to float around 0, which is a good sign. The second plot shows a similar

Feature	Category 1	Category 2	Category 3
forecast	1.000000	1.000000	1.000000
avg	0.476083	0.401405	0.377576
hour	0.633165	0.797530	0.636448
daylight	0.897537	0.614910	0.723164
movave_3	-0.146223	-0.215264	-0.138469
movave_7	-0.145918	-0.217873	-0.125057
movstd_3	-0.176330	-0.194209	-0.177661
movstd_7	-0.157247	-0.207805	-0.170879
movstd_30	-0.214293	-0.247470	-0.459833
movave_30	-0.114001	-0.222908	-0.439957
q90	-0.143794	-0.210506	-0.135834
business hour	0.224906	-0.151679	-0.019054
season	0.262519	0.280263	0.507304
q50	-0.147406	-0.195317	-0.117470
day	0.205025	0.270879	0.306666
weekend	0.058579	0.217230	0.258509
qtr	NaN	NaN	NaN
month	0.216172	0.244308	0.487091
q10	0.027524	-0.137596	0.221751

Table 4.4: Pearson Correlation for the predicted values by the ARIMAX model.

Category	ADF Stat	p-value	CV 1%	CV 5%	CV 10%	result
1	-15.378	0.000000	-3.432	-2.862	-2.567	stationary
2	-19.567	0.000000	-3.432	-2.862	-2.567	stationary
3	-12.778	0.000000	-3.432	-2.862	-2.567	stationary

Table 4.5: Augmented Dickey-Fuller test results for the residuals of the SARIMAX model, when applied on each of the 3 categories. CV stands for ”Critical Value”.

distribution of the errors to the non-seasonal model, which means that the errors are not normally distributed. There might be, again, some slight bias in the models. The third plot exposes similar trends of going off a normal distribution's line and in the same way, once again affirming the effect of the outliers on the model's predictions. Still, there are no significant autocorrelations in the residuals, which means the model did not leave any information in the residuals – it squeezed every bit of it out.

Table 4.5 shows that the series of residuals is stationary according to the Augmented Dickey-Fuller Test, which is a good sign that the errors do not correlate with one another in any way.

Secondly, the model's predictions are evaluated. Figure 4.4 shows the model's predictions compared to actual values on the testing set. How the split was done is described in chapter 3. It can be observed that the model is more willing than the non-seasonal

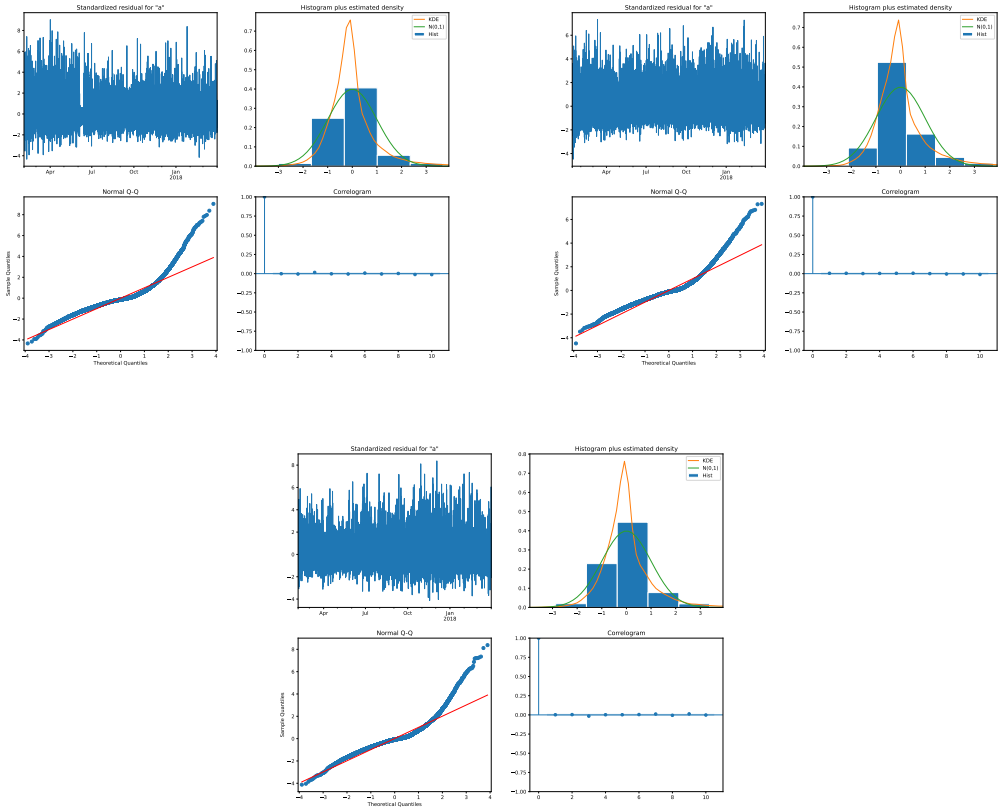

Figure 4.3: Diagnostics of predicted values by SARIMAX for each of the 3 categories, in order: category 1, category 2, category 3. Analysis of each diagnostic is presented in section 4.2.

one to adapt to extreme events, capturing the width of the shape a bit better. However, it fails to maintain the shape towards the end, for example in category 2, where there is a sharp decrease in consumption which the model does not seem to be able to account for. The gaps in data are still not captured. This is expected, since they do not occur often or at regular times. The prediction line also fails to follow the blue line towards the end of the testing period – that might have to do with the model's inability to predict on long periods of time, which is a known issue of ARIMA. There is also a suspicion that, in order to predict values deep in the testing period, it has used the previously predicted values as a guide, instead of using the actual registered values. This stems from the consistency in the shape, which seems not reflective of the actual trends but rather the predicted trends.

Table 4.6 sheds more light on the matter. It illustrates the correlation between the forecasted values and the other attributes. The predictions here are represented as attribute "forecast". Noticeable is that the correlation with the attribute for actual targets increased, compared to the non-seasonal model. That is definitely a sign that the shape is much better captured by the seasonal model. Correlations with the moving averages and standard deviations over the recent period, as well as percentiles, also increased and are now close to 0, sometimes a bit above 0, suggesting that the trend over time

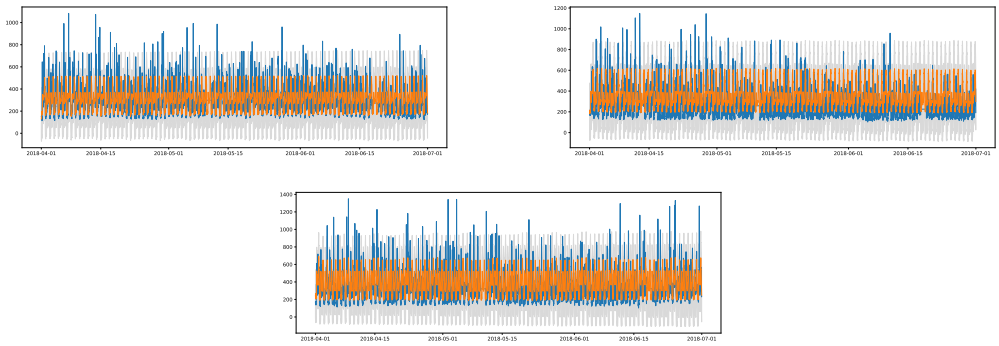

Figure 4.4: Results plot for each of the 3 categories, in order: category 1, category 2, category 3. The blue line represents actual consumption, orange line is the SARIMAX model's prediction, the grey hue is the 95% confidence interval.

is also better captured by the seasonal model. The highest correlations are still with the ”hour” and ”daylight” attributes, which, once again, seem to have been the most valuable features out of those fed to the model.

Table 4.7 shows a comparison between the SARIMAX and ARIMAX models based on the established metrics. One can see that results for the first category have slightly improved in all perspectives with the seasonal model, while for the other 2 categories, the results are way better for the non-seasonal one. It might be the case that for the last 2 categories the cycle has variable length overall, rather than being of fixed length and hence seasonal, while in the first category's data, seasonality was more present, hence the improvements in the results. It is interesting to note that for category 3, the seasonal model predicts slightly closer to the mean than the non-seasonal one, but is way further from the median. That is most likely an effect of the extreme events, which impacted the seasonal model greatly.

4.2.3 Comparing ARIMA against a Naive Baseline

In order to approve the results of the main model, it is important to establish a naive model against which the forecasting results of candidate models can be compared. For the IDEAL dataset, it has been shown that there are strong correlations between the hour mark and the target value, as well as between day of the week and target value. It has also been shown that weekly demand remains relatively consistent across the whole period. As such, we chose a seasonal naive forecast which imputes the value from the same interval the same day of the previous week. For example, the prediction for the interval 4:00PM - 4:30PM on Friday next week will be the value registered this week's Friday at 4PM - 4:30PM.

Table 4.8 presents a comparison between the results obtained by the naive model and the results obtained by the best ARIMA model for each category. The ARIMA models perform visibly better in all metrics.

Feature	Category 1	Category 2	Category 3
forecast	1.000000	1.000000	1.000000
avg	0.541458	0.535893	0.522818
hour	0.663633	0.782910	0.600683
daylight	0.677339	0.497677	0.660127
movave_3	0.000179	-0.009470	-0.007397
movave_7	-0.009364	-0.008343	0.001801
movstd_3	0.000471	-0.007574	-0.013915
movstd_7	-0.005706	-0.006056	0.001078
movstd_30	-0.009491	-0.009531	-0.008916
movave_30	-0.007129	-0.010039	-0.005262
q90	-0.007550	-0.007629	0.000507
business hour	0.162205	-0.094085	0.068430
season	0.011682	-0.001505	0.010297
q50	-0.011872	-0.008299	0.001686
day	0.058200	0.082198	0.097454
weekend	0.053554	0.121682	0.108211
qtr	NaN	Nan	NaN
month	0.009990	-0.000315	0.001223
q10	0.003918	0.001716	0.004379

Table 4.6: Pearson Correlation for the predicted values by the SARIMAX model.

	Category 1		**Category 2**		**Category 3**	
Metric	**SARIMAX**	**ARIMAX**	**SARIMAX**	**ARIMAX**	**SARIMAX**	**ARIMAX**
RMSE	108.14	109.57	147.55	120.06	151.14	152.06
MAE	77.86	84.48	121.02	91.27	115.47	102.89
MAPE	0.27	0.32	0.56	0.39	0.39	0.3

Table 4.7: Evaluation metrics for the SARIMAX and ARIMAX models for each category.

	Category 1		**Category 2**		**Category 3**	
Metric	**Naive**	**ARIMA**	**Naive**	**ARIMA**	**Naive**	**ARIMA**
RMSE	131.65	108.14	146.79	120.06	168.67	152.06
MAE	92.05	77.86	102.06	91.27	117.02	102.89
MAPE	0.31	0.27	0.39	0.39	0.36	0.3

Table 4.8: Evaluation metrics for the naive model and the best ARIMA model for each category.

Chapter 5

Conclusion

5.1 Software application

The software application, which lets the user know the predicted electricity consumption, is available in the project materials. Just run ”python main.py” and the user interface shall load. Make sure you have the following libraries installed: kivy, pickle, joblib, pandas, matplotlib, sklearn, statsmodels, numpy, os, pathlib, sys, functools, time and math.

The user interface is quite simple. Once the app runs, a page will open up containing a form, as illustrated in Figure 5.1. The user is able to input the type of house, the number of people and when it was built, then click submit and wait for the compiler to quickly load the model and show the graph of the predicted consumption, as well as the average daily root mean squared error, mean absolute error and mean absolute percentage error. To predict these, the app makes use of pre-saved models. It selects the model and the dataset to apply it on based on the user input. Note that due to dimension reasons, the models loaded in the app are the non-seasonal ones.

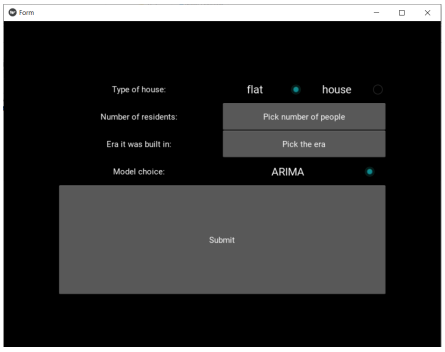

Figure 5.1: Screenshot of the application’s UI interface.

5.2 Discussion

This paper presented a way of categorizing households, aggregating the homes in each category and predicting each category's average electricity consumption for every half-hour. The IDEAL dataset proved very fruitful for the task, delivering enough data for an in-depth analysis.

Exploring the data has delivered information on multiple fronts. It has shown the difficulties of aggregating data and how much variation there is in people's behaviour, even if the households are part of the same categories. Then, it proceeded to discuss the form that aggregated data takes: it looked at seasonality, trend, variation, volatility, density and correlations. Some interesting facts are summarized below:

- There is no apparent trend in the aggregated use of electricity consumption per half-hour. That is expected at household, since people use electricity based on needs, which are on and off and not every day in the same half-hour. The absence of a trend at group level, however, underlines the influence of external factors, like pricing or weather.
- There is no visible seasonality, at least not detectable by the Augmented Dicky-Fuller test. However, the autocorrelations unveiled a certain seasonality pattern from one day to another, which was later confirmed for one of the categories by the better results of the seasonal model compared to the non-seasonal one. At household level, one might expect seasonality given that people have some routine in their lives, but at subsector level, it was not as obvious.
- There is a lot of variation and volatility in electricity data at 30-minute level, even at group level. However, the density plot has shown that generally the aggregated usage tends to resume around a certain mean, and there is not much variation around the mean.
- Looking at correlations and dependencies in the data was insightful. Besides unveiling a certain seasonality pattern, it also showed small tendencies in electricity consumption, from one half-hour to another.

Of course, the amount of data at disposal was limited. That lead to forming categories and considering the averages of 5 to 10 households for each category, which is not a big enough sample to draw any relevant conclusions from at a high level. Besides, data was available for a bit more than 520 days at maximum for a category – that is, approximately 2 months less than a year and a half.

However, the amount of available data was not the only limiting factor. The model's capabilities were another one. While SARIMAX has proved to be an effective linear model that can grasp the linear trend and seasonality of time series data in many cases [36], the assumption of linearity in itself is limiting. It makes the model vulnerable to sudden changes and unexpected behaviour, which is most certainly the case with people's behaviours and their way of utilising electricity. Besides, SARIMAX is computationally expensive, making it extremely difficult to compute for datasets where correlations are found at a high number of lags apart.

Another limitation stems from the limited amount of accuracy one can get when predicting electricity consumption. There are a lot more factors that need to be taken into account, and for some of these factors, there is not enough data just yet.

5.3 Further work

The work in here mostly lays ground and sets some expectations for analysis on bigger and more sophisticated groups of households. There are, of course, some other 5 categories yet to be explored, which in here was not possible due to space limitations. There is enough room for improvement and this section will underline possible routes of achieving improvements.

An obvious one would include trying to reduce the number of limitations that exist for the present paper in order to proceed with further, more in-depth, higher-level analysis. The most relevant would be addressing the accuracy issue, since until that is done, any model optimization would only squeeze so much out of the data at hand. As discussed in chapter 2, it certainly is the case that people's behaviours should be studied and be taken into account for a more complete, effective conclusion. One way of doing this is obviously by organising surveys, which has already been done in certain regions, such as north-east of Scotland [7]. Another one would be by studying causality – that is, analysing possible roots of causes for the different patterns in the data, like it was done in Bangladesh with GDP [21].

Another way would be to build further ground for more complete analysis on bigger datasets. There has been analysis done on splitting houses into categories [16] [12], analysis on consumption patterns and trends at individual level and sometimes at higher levels, like national level [25] [19], even if not at the same granularity. Consideration could go into increasing accessibility of optimization methods in terms of space complexity, i.e. methods which optimize how much storage an algorithm needs, given the size of the dataset it has to process. Some solutions already exist, such as genetic algorithms and fruit fly optimization, and they have been applied in analysing electricity consumption too [24] [8]. However, most of these techniques still need to be implemented by the developer, and so they are not easily accessible to the average scientist outside the field of computer science.

5.4 Plan for MInf Part 2

Part 2 will focus on bringing meaningfulness to the analysis. For most purposes in managing the electricity grid, there is not much interest in managing the individual household, but rather groups of households like, for example, neighborhood areas. In this paper, it was visible that aggregating data from only a few houses brought in huge standard deviation and then the accuracy of the model was not very good, due to extreme events happening ever so often. This is expected for a low number of houses, since people's behaviour largely differs from one household to another, and extreme events take a visible toll on the aggregate for the consumption during one half-hour. However, there is an argument to be made that combining data from a

large number of houses would lead to smaller standard deviation and to sharper, better-defined, aggregate consumption trends. This would, in turn, lead to better prediction accuracy and to more easily definable distribution of residuals.

This hypothesis, however, needs to be tested. Hence, part 2 will look at exactly that, making use of the data collected at half-hourly rate from 13300 houses in the UK by the Smart Energy Research Lab, with the help of smart meters [34].

Bibliography

[1] Decarbonisation pathways for cooling and heating. https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/V042955/1.

[2] Ideal dataset. https://datashare.ed.ac.uk/handle/10283/3647.

[3] Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

[4] Musaed Alhussein, Khursheed Aurangzeb, and Syed Irtaza Haider. Hybrid cnn-lstm model for short-term individual household load forecasting. *IEEE Access*, 8:180544–180557, 2020.

[5] F.M. Andersen, P.A. Gunkel, H.K. Jacobsen, and L. Kitzing. Residential electricity consumption and household characteristics: An econometric analysis of danish smart-meter data. *Energy Economics*, 100:105341, 2021.

[6] Khursheed Aurangzeb, Musaed Alhussein, Kumail Javaid, and Syed Irtaza Haider. A pyramid-cnn based deep learning model for power load forecasting of similar-profile energy customers based on clustering. *IEEE Access*, 9:14992–15003, 2021.

[7] Greta Barnicoat and Mike Danson. The ageing population and smart metering: A field study of householders' attitudes and behaviours towards energy use in scotland. *Energy Research Social Science*, 9:107–115, 2015. Special Issue on Smart Grids and the Social Sciences.

[8] Guohua Cao and Lijuan Wu. Support vector regression with fruit fly optimization algorithm for seasonal electricity consumption forecasting. *Energy*, 115:734–745, 2016.

[9] Tony Craig, J. Gary Polhill, Ian Dent, Carlos Galan-Diaz, and Simon Heslop. The north east scotland energy monitoring project: Exploring relationships between household occupants and energy usage. *Energy and Buildings*, 75:493–503, 2014.

[10] Neil Darragh. AI Forecasting of Dissagregated Demand.

[11] Susan Dean and Barbara Illowsky. Descriptive statistics: skewness and the mean, median, and mode. *Connexions website*, 2018.

[12] Carolina Madeira R. do Carmo and Toke Haunstrup Christensen. Cluster analysis of residential heat load profiles and the role of technical and household characteristics. *Energy and Buildings*, 125:171–180, 2016.

[13] A. Druckman and T. Jackson. Household energy consumption in the uk: A highly geographically and socio-economically disaggregated model. *Energy Policy*, 36(8):3177–3192, 2008.

[14] Ottmar Edenhofer. *Climate change 2014: mitigation of climate change*, volume 3. Cambridge University Press, 2015.

[15] Brian S Everitt and Anders Skrondal. The cambridge dictionary of statistics. 2010.

[16] Hideitsu Hino, Haoyang Shen, Noboru Murata, Shinji Wakao, and Yasuhiro Hayashi. A versatile clustering method for electricity consumption pattern analysis in households. *IEEE Transactions on Smart Grid*, 4(2):1048–1057, 2013.

[17] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[18] Weicong Kong, Zhao Yang Dong, Youwei Jia, David J Hill, Yan Xu, and Yuan Zhang. Short-term residential load forecasting based on lstm recurrent neural network. *IEEE Transactions on Smart Grid*, 10(1):841–851, 2017.

[19] Shuyu Li and Rongrong Li. Comparison of forecasting energy consumption in shandong, china using the arima model, gm model, and arima-gm model. *Sustainability*, 9(7), 2017.

[20] Javier López Prol and Wolf-Peter Schill. The economics of variable renewable energy and electricity storage. *Annual Review of Resource Economics*, 13:443–467, 2021.

[21] Pallab Mozumder and Achla Marathe. Causality relationship between electricity consumption and gdp in bangladesh. *Energy policy*, 35(1):395–402, 2007.

[22] Economic Outlook OECD. Sources and methods, 1987.

[23] Department of Energy and Climate Change (DECC). Energy consumption in the united kingdom, 2009.

[24] Harun Kemal Ozturk, Halim Ceylan, Olcay Ersel Canyurt, and Arif Hepbasli. Electricity estimation using genetic algorithm approach: a case study of turkey. *Energy*, 30(7):1003–1012, 2005.

[25] Suat Ozturk and Feride Ozturk. Forecasting energy consumption of turkey by arima model. *Journal of Asian Scientific Research*, 8(2):52–60, 2018.

[26] Alan Pankratz. *Forecasting with univariate Box-Jenkins models: Concepts and cases*. John Wiley & Sons, 2009.

[27] KARL PEARSON. "DAS FEHLERGESETZ UND SEINE VERALLGEMEINER-UNGEN DURCH FECHNER UND PEARSON*." A REJOINDER. *Biometrika*, 4(1-2):169–212, 06 1905.

[28] Martin Pullinger, Jonathan Kilgour, Nigel Goddard, Niklas Berliner, Lynda Webb, Myroslava Dzikovska, Heather Lovell, Janek Mann, Charles Sutton, Janette Webb, et al. The ideal household energy dataset, electricity, gas, contextual sensor data and survey data for 255 uk homes. *Scientific Data*, 8(1):1–18, 2021.

[29] Meysam Qadrdan, Meng Cheng, Jianzhong Wu, and Nick Jenkins. Benefits of demand-side response in combined gas and electricity networks. *Applied energy*, 192:360–369, 2017.

[30] Debneil Saha Roy. Household level electricity load forecasting using echo state network. In *2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, pages 1–6. IEEE, 2020.

[31] Andreas Schmidt, Ana Ivanova, and Mike S. Schäfer. Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. *Global Environmental Change*, 23(5):1233–1248, 2013.

[32] Matthew R. Sisco, Silvia Pianta, Elke U. Weber, and Valentina Bosetti. Global climate marches sharply raise attention to climate change: Analysis of climate search behavior in 46 countries. *Journal of Environmental Psychology*, 75:101596, 2021.

[33] Abdul Wahab, Muhammad Anas Tahir, Naveed Iqbal, Faisal Shafait, and Syed Muhammad Raza Kazmi. Short-term load forecasting using bi-directional sequential models and feature engineering for small datasets. *arXiv preprint arXiv:2011.14137*, 2020.

[34] Ellen Webborn, Jessica Few, Eoghan McKenna, Simon Elam, Martin Pullinger, Ben Anderson, David Shipworth, and Tadj Oreszczyn. The serl observatory dataset: Longitudinal smart meter electricity and gas data, survey, epc and climate data for over 13,000 households in great britain. *Energies*, 14(21), 2021.

[35] Robert Fraser Williamson, Andrew Sudmant, Andy Gouldson, and Jamie Brogan. A net-zero carbon roadmap for edinburgh. *Edinburgh Climate Commission/Place-Based Climate Action Network*, 2020.

[36] Xingyu Zhang, Yuanyuan Liu, Min Yang, Tao Zhang, Alistair A Young, and Xiaosong Li. Comparative study of four time series methods in forecasting typhoid fever incidence in china. *PloS one*, 8(5):e63116, 2013.