VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
DATA SCIENCE
MASTER'S STUDY PROGRAMME

Electricity customer segmentation and price sensitivity identification from electricity load pattern

Elektros vartotojų segmentavimas ir ir jautrumo kainai identifikavimas remiantis elektros suvartojimo duomenimis

Master's thesis

Author: Domas Meilūnas
VU email address: domas.meilunas@mif.stud.vu.lt
Supervisor: Dr. Jurgita Markevičiūtė

Vilnius
2024

Contents

Abstract

With the coming of smart grid comes new analytical possibilities and challenges for companies. Increased quantities of data open new possibilities for customer analytics like tracking power consumption behaviors of customers or customer classification according to their electricity consumption patterns or identification of pricing effect on electricity usage. Understanding the impact of pricing structures on consumer energy usage at different times of the day can be instrumental for energy providers and policymakers in understanding and designing pricing strategies that not only encourage more efficient energy use but also effectively manage the demand on the energy grid.

This paper investigates possibility to identify price sensitivity of user just from their electricity usage habits and construct segmentation model to classify them into one of load patterns from historic electricity consumption data.

Keywords: Functional Analysis, Smart metering, Clustering, Utilities, Price Sensitivity, FCM, CART

Santrauka

Su išmaniųjų skaitiklių atėjimu atsiranda naujų analitinių galimybių ir iššūkių energetikos įmonėms. Didėjančios duomenų apimtys atveria naujas galimybes duomenų analizei, pavyzdžiui, stebėti klientų elektros energijos suvartojimo įpročius, klasifikuoti tuos pačius klientus pagal jų elektros energijos suvartojimo įpročius, ar identifikuot kainos pokyčio poveikį elektros energijos naudojimui. Šios informacijos turėjimas gali būti naudingas tiek nepriklausomiems tiekėjams, formuojant kainodaros strategijas, tiek rinkos reguliatoriams formuojant principus ir gaires skatinant efektyvesnį energijos naudojimą ir valdymą.

Šiame darbe tiriamos galimybės nustatyti vartotojo kainų jautrumą tik iš jų elektros energijos suvartojimo istoriių duomenų ir bandoma sukurti segmentavimo modelį, kuris galėtų vartotojus suklasifikuoti į grupes reiantis elektros energijos vartojimo įpročiais.

Raktažodžiai: Funkcinių duomenų analizė, Išmanieji skaitikliai, Energetika, Klasterizavimas, jautrumas kainai, FCM, CART

1 Introduction

The energy sector in Lithuania is currently undergoing significant transformation, which, should bring couple of changes to end users. The most substantial change is the liberalization of the market which allows the entry of private companies into the market. This change breaks up monopolized energy supply and empowers consumers with the choice to select their electricity provider. As the project already reached third stage according to Lithuania's Energy Agency as of end of year over 1.2m users were subject to market liberalization and had picked energy provider [1]

Concurrently, the country is advancing with the installation of a smart grid, which will bring new possibilities and challenges for data analysis. Newest press release by countries distribution system operators (DSO) stated that over 700 thousand smart meters were installed throughout the country and by 2026 this number should reach 1.2 million [12]. With market liberalization and smart grid on the way this creates new issues to energy providers, with abundance of data and customer, with freedom to move between providers this transformation necessitates that companies develop methods to swiftly classify and evaluate customer data.

In the wake of these developments, personalized pricing strategies can become a pivotal approach. This strategy, which moves away from the conventional one-size-fits-all model, accounts for the unique energy needs and consumption patterns of individual customers. By adopting these personalized strategies, utilities can foster energy-saving behaviors, redistribute consumption to off-peak periods to ease grid stress, and enhance overall energy efficiency. This customer-centric methodology not only improves customer satisfaction and engagement but also supports sustainable energy consumption. The latter is especially critical in the context of global energy challenges and the imperative transition toward a more sustainable energy economy.

As it was mentioned in [5] successful implementation of personalized dynamic pricing hinges on the development of effective customer segmentation and pricing strategies for retail electricity consumers.Such strategies are integral for aligning pricing models with the diverse consumption patterns and preferences of individual users, ultimately driving more efficient and equitable energy use. One of the primary obstacles for a deep understanding of customer is gauging their price sensitivity and willingness to pay. In the same paper [5] it was referenced that depending on the market there might be customers who would be "willing" to pay up to 1.5 times more. These elements are crucial for shaping pricing strategies that consumers are likely to accept and adhere to, ensuring that dynamic pricing fulfills its potential in managing demand and promoting, for example, sustainability. Utilities, therefore, must employ innovative approaches to infer these critical factors, utilizing advanced data analytics tools and customer engagement techniques to bridge the knowledge gap.

However, investigation of price sensitivity identification from load pattern proved to be futile. Academic papers were mainly focused more on identification of macro level price elasticity,

or focused on use of quantitative data. For example in [7] household-level panel data was used and found strong evidence that consumers respond to average rather than marginal price. Other study in Denmark [13] used panel data from Danish Building and Housing Register and investigated price elasticity of residential district heating demand and found significant evidence that price elasticity varies across household groups. On the other hand there are more research done on broader price elasticity topic. [11] conducted research on French price elasticity of electricity expenditure of private households, or [2] did very similar analysis of price elasticity for residential electricity consumers in Poland.

In contrast, methods to recognise and classify load profile have been researched in numerous studies over the last 20 years. Number of papers were reviewed which took different approaches for the same problem. The researches in [6] and [4] used a self organizing mapping neural network to obtain the possible number of load pattern clusters, which later were used in other supervised methods. In [10] a bit more practical implementation was described, a Fuzzy Cmeans (FCM) algorithm was used for load pattern recognition and Classification And Regression Tree (CART) together with Load Characteristics Index (LCI), introduced in [3], was used for Load profile prediction. Each of the above method has its own limitation. For example majority of the reviewed research papers indicated different algorithms that can be used for load profile classification, however, at the end focused on same clustering method like k-means [6]. Only two of the reviewed papers [14] and [15] did comparison of few clustering algorithms. In [14] number of different algorithms were examined, however, it did not managed to provide any detailed summary of the results. On the other hand [15] compared only two algorithms, k-means and FCM and did not find any significant differences between the received results. Some newer papers provided more insights in using wider models such as [8] where they used spectral clustering for feature-based pricing.

Due to all this complexity and raising need for companies to be able to classify and evaluate customer on the spot, for example in digital channels to offer price quote, in most cases companies can rely only on customer's historic energy consumption patterns. With that in mind this paper investigates if any observable link between consumption and price sensitivity can be identified and proposes two-way clustering method to better understand users' electricity consumption behavior. By applying this method, the research aims to identify resembling consumption patterns and discern customers who would be subject to higher price sensitivity. The goal of this research is not only to identify if such patterns exist but also to provide a robust framework through which utilities can leverage consumption data for strategic customer segmentation. This could have significant implications for the future of dynamic pricing in the retail electricity sector, facilitating a more data-driven, customer-centric approach to energy management.

2 Data and methodology

2.1 Data set

The data-set used in research originates from the Low Carbon London initiative [[9]], a project spearheaded by UK Power Networks. This project, which ran from November 2011 until February 2014, gathered detailed energy consumption data from a sample of 5,567 households within London. The comprehensive data-set contains around 167 million observations, each recorded at 30-minute intervals. Due to constraints related to data processing and performance, a smaller subset consisting of 30 million observations was utilized for in-depth analysis.

The study primarily focused on households participating in a Dynamic Time of Use (ToU) electricity pricing scheme. Participants in this scheme were informed a day in advance about the varying electricity prices, tariffs were categorized into three levels: High, priced at 67.20 pence per kilowatt-hour (p/kWh); Low, at 3.99 p/kWh; and Normal, at 11.76 p/kWh. In contrast, households not enrolled in the Time of Use program were charged a consistent flat rate of 14.228 p/kWh, regardless of their time of consumption.

2.2 Pre-processing

Due to the sophisticated nature of its collection methods the data-set required pre-processing.

The initial phase of pre-processing involves a cleansing of the data to rectify any discrepancies that may have arisen from malfunctions in the smart meters or from anomalies in data transmission across the network. Notably, there were identified several instances where meters unexpectedly reported 'null' values. These missing values are presumed to be outliers, potentially resulting from technical issues such as the meter detecting unexpected voltage fluctuations or there being a network issues etc. Because there were no information provided on the actual smart meters used by household there was no means to validate possible causes of missing values. To ensure the data-set's accuracy and reliability, these null values were removed

Data-set can be expressed like this:

$$U = [U_1, U_2, ..., U_i, ..., U_N]^T \tag{1}$$

$$U_i = [u_{i1}, u_{i2}, ..., u_{ij}, ..., u_{i48}] \tag{2}$$

Here U - observed household, U_i - number of observations at household and N - number of observed households.

Following the cleansing stage, the data-set is subject to a observation reduction. Given the extensive amount of data collected over multiple days, it was necessary to condense this information into a more manageable form. This was achieved by calculating the average daily energy consumption for each household per time observation. Specifically, this involved taking

the mean of all the readings for each household at each 30-minute interval throughout the day.

2.3 Data Analysis and Observations

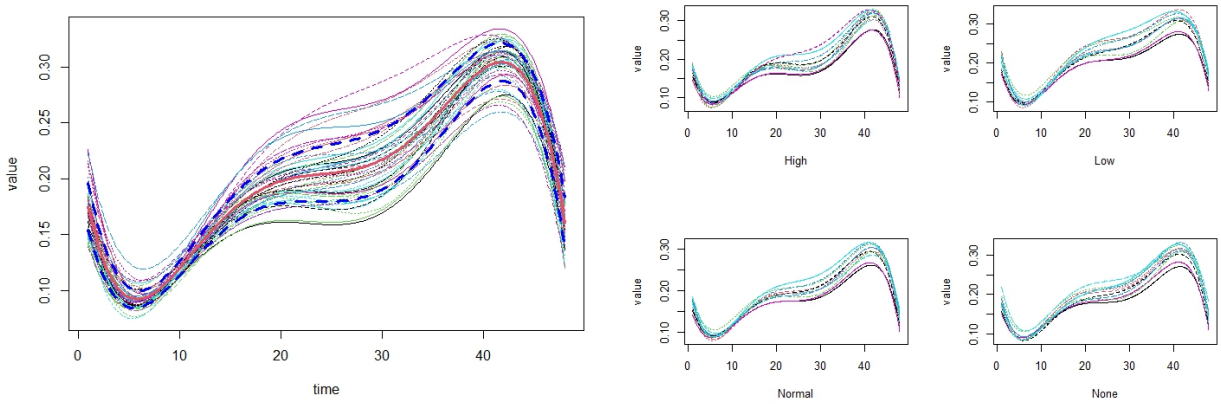

Figure 1: Mean analysis

Each of observed households were randomly assigned to one of 12 groups and for each group average consumption for each tariff category were calculated resulting in 48 observations. Then B-Spline smoothing with order 6 and smoothing parameter 0.32 were applied to convert data into functional data. As a result the Root Mean Square (RMS) variated between 0.01 and 0.02. Further analysis of the data provided additional insights that were used further in research. Mean and variation revealed that the highest variations in energy usage tended to occur around the middle of the day. Interestingly, this pattern of variation was observed to varying degrees across each of the tariff categories. The distinct trends and variations uncovered through this analysis are visually represented in Figure 1.

The derivative analysis conducted as part of this research served to reaffirm the trends observed in the mean consumption graphs - Figure 1.

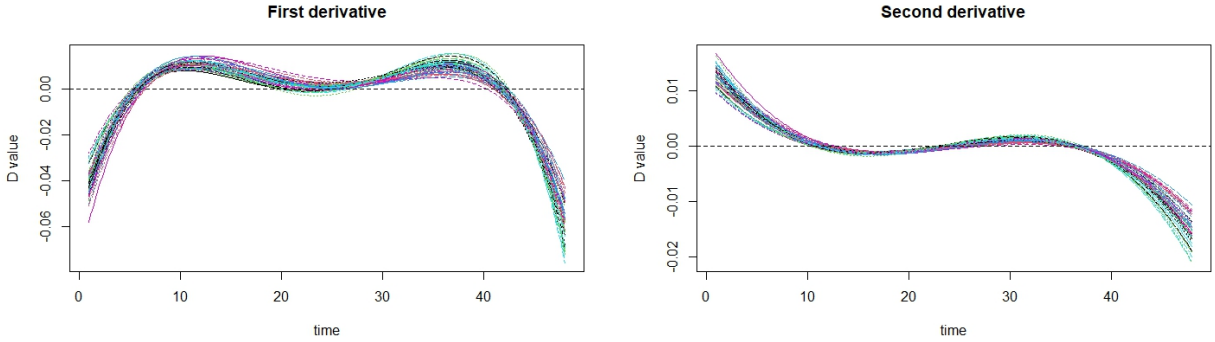

Figure 2: Derivative tests

The findings from the derivative test revealed two major turning points in energy usage patterns: one at the start of the day and another at the end. The start of the day is characterized by a marked increase in energy consumption which is then followed by a period of relatively stable usage throughout midday, indicating a plateau in consumption levels. As the day progresses towards the night, a significant drop in energy consumption is observed. Following observation is to be expected following general cycle of humans' activity through the day.

Interestingly Principal Component Analysis (Figure 3) showed that majority of the variability in household energy consumption occurs primarily in the early morning hours and disappears almost entirely by mid-day. This pattern suggests that morning energy use is a key determining household's overall daily energy consumption. Additionally, a Varimax Rotation provided the same results concluding that the power consumption at the start has biggest impact explaining general tendency of household's average power consumption throughout the day.

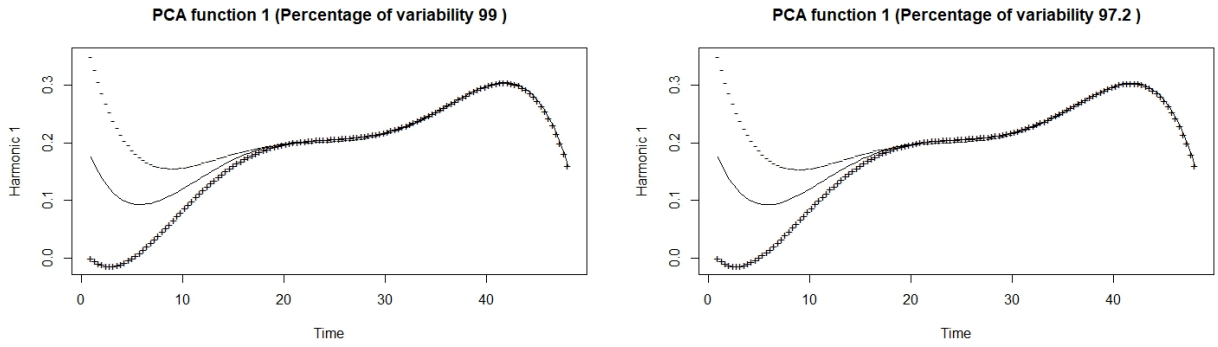

Figure 3: Principal Component Analysis (PCA)

These insights are particularly valuable as they highlight the morning period as a critical target for energy-saving initiatives and can guide energy providers in designing more effective energy management strategies.

2.4 Price Sensitivity analysis

As stated before the prior knowledge of price sensitivity among consumers offers additional strategic advantages to energy providers. By comprehending how different customer segments respond to price changes, energy companies can tailor their pricing strategies more effectively. This knowledge is crucial in designing dynamic pricing models, such as Time of Use (ToU) tariffs, which can influence consumer behavior and lead to more efficient energy usage patterns. Purpose of all these analysis was to understand if knowing electricity prices in advance has impact on consumer behavior and their power consumption in general.

To do so the following hypothesis was derived to be tested:

$$H_0 : \mu_1(t) = \mu_2(t) = \mu_3(t) = \mu_4(t) \tag{3}$$

$$H_1 : \text{The means are not all equal} \tag{4}$$

Here each μ_i represent mean consumption of each tariff category

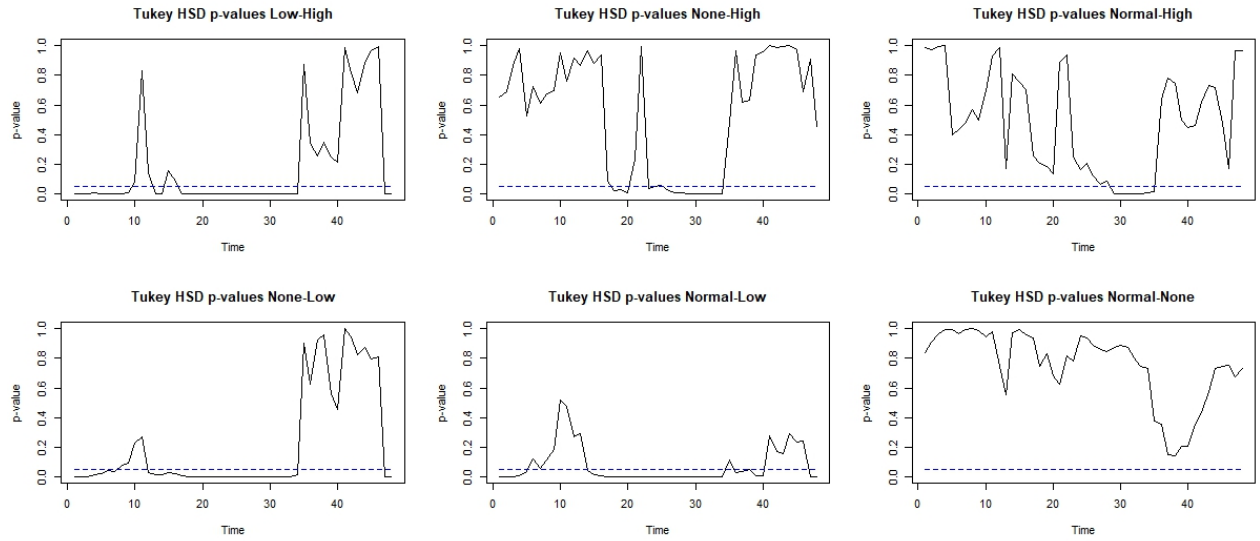

Figure 4: fANOVA group comparison

For the purpose of analyzing variances across the means of different pricing indication groups, a functional ANOVA model was selected. Interestingly, the most pronounced differences in

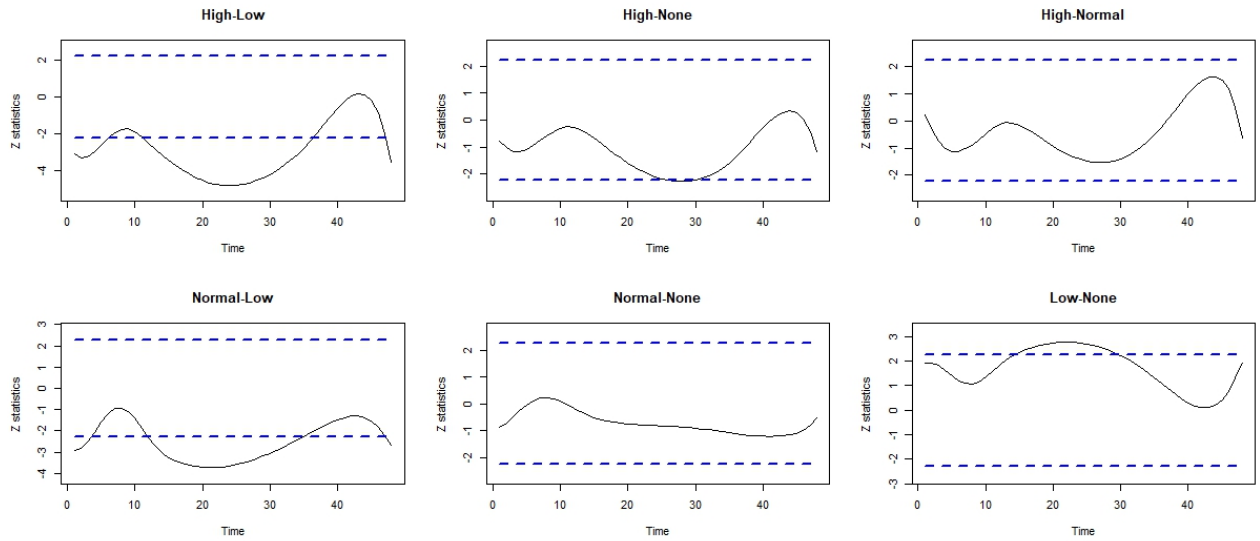

Figure 5: Two sample pointwise test results

variance were observed at two key points of the day. The first significant variation occurs at the very start/end of the day, suggesting that the initial energy usage behaviors vary considerably under high/normal and low tariff conditions. The second notable observation of variance occurs over lunchtime up to the end of working day. (Figure 4)

To validate the results obtained from the functional ANOVA analysis, two-sample pointwise tests were conducted. Two-sample pointwise test, like fANOVA, designed to compare two groups at each point in time, providing a detailed examination of the differences in variance between them. The pointwise testing reconfirmed the findings of the fANOVA, specifically regarding the variance between the High/Low and Normal/Low tariff groups (Figure 5). By applying these pointwise tests, the study was able to reinforce the earlier observations that significant differences in energy consumption patterns exist between these groups and rule out uncertain groups (High/None, High/Normal). This additional layer of analysis strengthens the confidence in the original fANOVA results, providing robust evidence that the price changes significantly impacts household energy usage patterns.

3 Clustering and Results

For the purpose of identifying distinct usage patterns within the data-set, Fuzzy C-Means clustering algorithm was employed. This method is particularly effective in partitioning the observations into clusters, where each observation is grouped based on its proximity to the nearest cluster mean. Unlike traditional clustering methods which assign each observation to a single cluster, Fuzzy C-Means allows for a degree of uncertainty, enabling each observation to belong to multiple clusters to varying extents. This soft clustering technique is especially useful in scenarios where data points are not distinctly separable or when there is an overlap in characteristics among different groups. In the classification aspect of the analysis, the CART (Classification and Regression Tree) algorithm was utilized. CART is a versatile decision tree technique that can be used for both classification and regression tasks. In this context, it serves to classify different households based on their energy usage patterns. The algorithm works by recursively dividing the data-set into subsets based on the attributes that most effectively differentiate the observations. This results in a tree-like model of decisions, which can be particularly insightful for understanding the factors that influence different energy consumption behaviors.

Based on findings of data analysis the following characteristics were introduced for classification and regression trees (Table 3). The first characteristic, Load Factor, measures the

Description	Time	Index	Definition
Load factor	0:00-24:00	a1	$=P_{avgperiod}/P_{max}$
	0:00-24:00	a2	$=P_{min}/P_{max}$
Day start	0:00-6:00	a3	$=P_{avgperiod}/P_{max}$
Peak usage	6:00-9:00, 17:00-21:00	a4	$=P_{avgperiod}/P_{avg}$
Off-peak usage	9:00-17:00, 21:00-24:00	a5	$=P_{avgperiod}/P_{avg}$

Table 1: Characteristics index

efficiency of electrical energy usage. A high load factor suggests more efficient and stable use of the electric system. This metric is particularly useful in assessing how consistently electricity is consumed. Range Ratio, quantifies the difference between the highest and lowest usage points and provides insight into the variability of consumption, indicating the extent to which consumption levels fluctuate. Additionally, Day Start usage was selected based on findings from the Principal Component Analysis (PCA), focusing on fluctuations in electricity usage during the early morning hours. This aspect helps in understanding the initiation of daily consumption patterns. Lastly, the model considers Peak and Off-Peak Usage, examining fluctuations during times of high and low electricity demand. This helps in understanding how consumption varies

with respect to different times of the day, which is crucial for identifying usage patterns and potential areas for efficiency improvements.

The integration of Fuzzy C-Means for clustering and CART (Classification and Regression Tree) for classification, as applied in this research, has demonstrated noteworthy efficiency in terms of performance. This effectiveness is particularly significant as it opens up promising opportunities for these methods to be implemented in practical, real-world applications.

The combination of these two methods not only enhances the accuracy of the analysis but also ensures a level of computational efficiency that makes them viable for large-scale application. This could have significant implications in areas such as energy management, policy-making, and customer segmentation in the energy sector. The potential for these techniques to contribute to more efficient and sustainable energy use in real-life settings is substantial, marking an exciting development in the field of energy data analysis.

3.1 Load Pattern Clustering

3.1.1 Data preparation

Before proceeding with clustering data had to be normalized. This process is essential for enabling accurate and meaningful comparisons across different households. Normalization adjusts the data to a uniform scale, thereby facilitating a more straightforward and effective analysis ensuring that the data-set is both manageable and primed for detailed analytical examination, maintaining its integrity and enhancing its value for research and insights.

Simple normalization method, described in previous research [10], was chosen (5) - to normalize data each observation from U_i is divided by max U_i of each U.

Thus normalized observations then can be expressed as follows:

$$D_i = U_i / \max(U_i) \tag{5}$$

where the max U_i is the highest household consumption value observed throughout the day.

3.1.2 Clustering and Results

The determination of the optimal number of clusters for analysis presented somewhat of a challenge, as different methods yielded varying results (Figure 6). The Elbow method, a popular method used to determine the number of clusters in a data-set, suggested the use of two clusters. In contrast, the Gap Statistic, another widely used technique recommended to use up to ten clusters. This discrepancy between the Elbow method and Gap Statistic results highlights a common issue in cluster analysis - the lack of a one-size-fits-all approach to determining the optimal number of clusters. The Elbow method, which involves plotting the explained variance against the number of clusters and looking for a point where diminishing returns are offset by

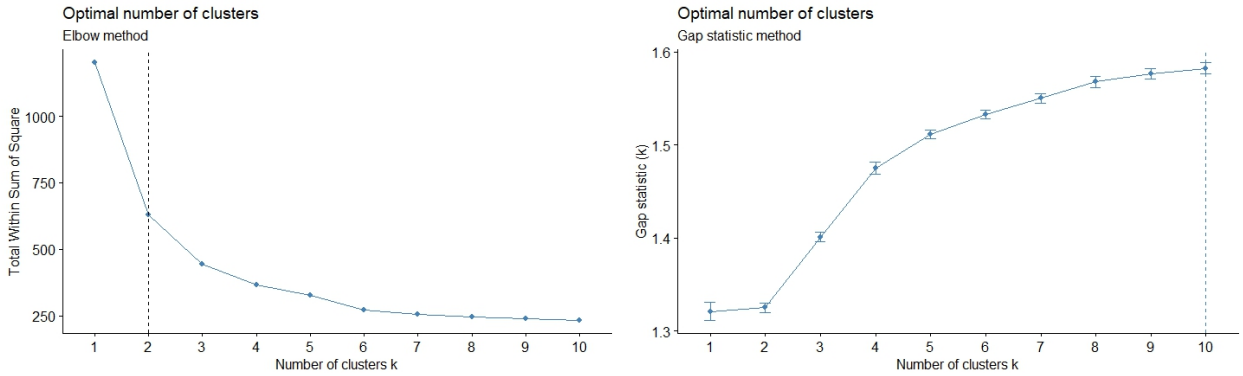

Figure 6: Cluster analysis

additional cost, suggested a simpler model with two clusters. However, the Gap Statistic, which compares the total within intra-cluster variation for different numbers of clusters with their expected values under null reference distribution of the data, indicated a more complex model with ten distinct groups. After validating all in between options it was decided to proceed further analysis with three clusters as it provided reasonable amount of differentiation without adding unnecessary complexity.

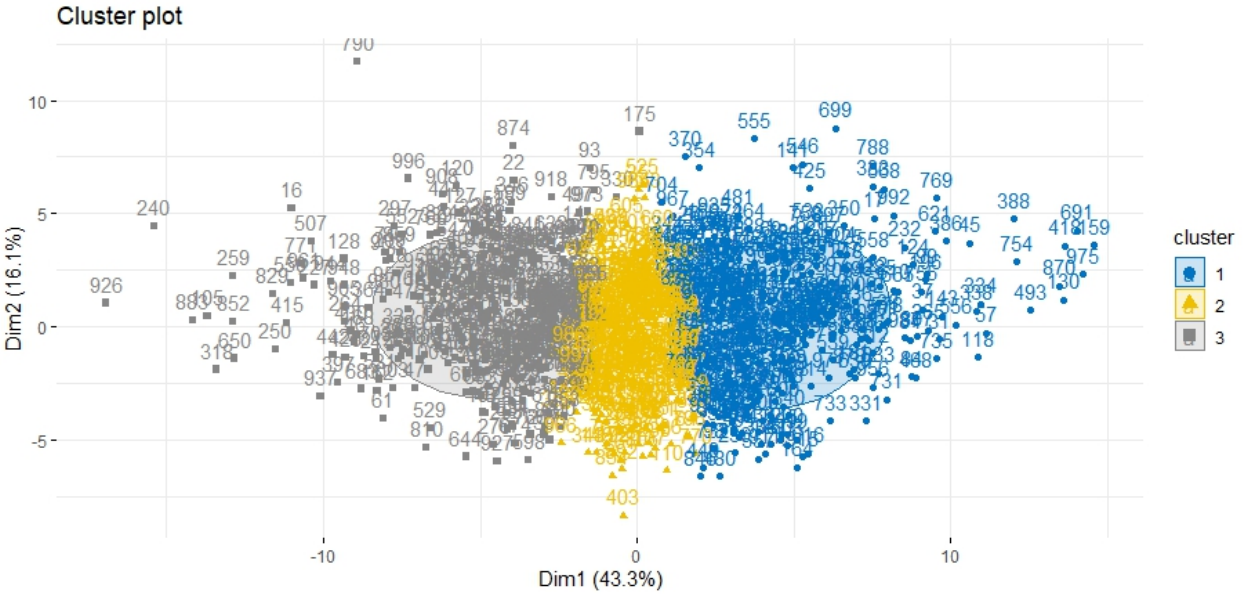

Figure 7: FCM clusters.

Fuzzy C-means cluster center plot provided promising results, as it can be seen from Figure 7 there is clear separation between three clusters. The lack of significant overlapping regions in the cluster plot is a key observation. When using Fuzzy C-Means clustering, some degree of overlap is common due to the algorithm's nature of allowing data points to belong to multiple clusters to varying extent. However, the clear separation observed in this case suggests that the energy consumption patterns of households in each cluster are distinct. This distinction is

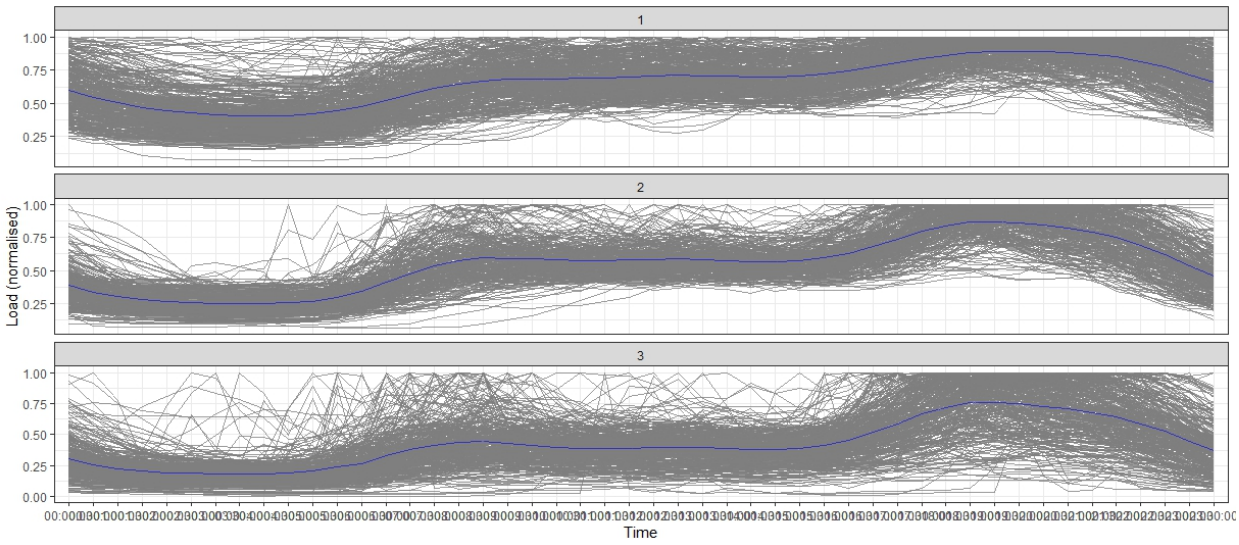

Figure 8: Clusters means.

essential for accurately categorizing households based on their energy usage behaviors.

The cluster mean plot, as shown in Figure 8, offers further insight into the distinctiveness of the three clusters identified by the Fuzzy C-Means algorithm. Despite all three clusters exhibiting somewhat similar usage patterns, they each possess distinguishable mean values. This observation aligns with the findings from the cluster center plot, reinforcing the conclusion that each cluster is distinguishable despite similarity in usage patterns across the clusters that can be explained by commonalities in how energy is consumed by household consumers. However, the differences in the mean values of each cluster highlight variations in the intensity or volume of energy usage among them. Proceeding with CART model values from characteristics index

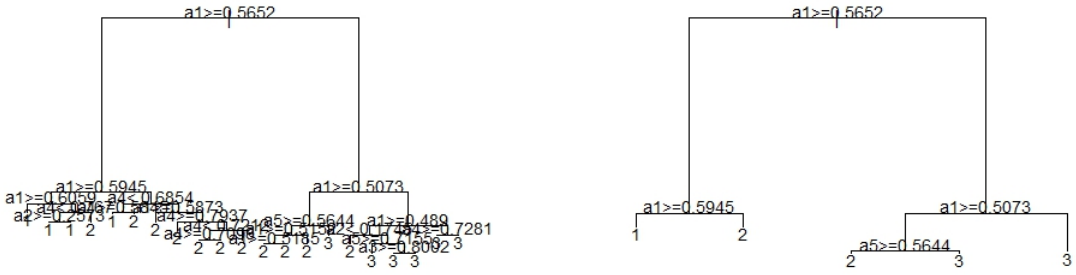

Figure 9: Classification and Regression Tree

(Figure 3) were used as inputs for classification tree. Classification tree was pruned with a complexity parameter of 0.0055. This pruning process reduces the tree's complexity without

significantly compromising its predictive power. The resulting Figure 9 indicates that only two characteristics from the index were significant enough to be used in the pruned classification tree. The use of a small number of characteristics suggests that these factors are highly indicative of the classification categories and that additional data points may not significantly improve the model's performance. The performance of the CART model was evaluated by com-

	1	2	3	%
1	347	28	0	92.5%
2	8	261	25	88.8%
3	0	5	337	98.5%

Table 2: Classification Results

paring the predicted results from the classification tree to the actual clustering results for each household. Based on the results seen in Table 2 prediction accuracy for clusters 1 and 3 was above 92% and cluster 2 was just a bit behind with almost 89%. In overall the CART model with Characteristics Index performance in classifying households into the correct groups based on their energy consumption patterns can be seen as highly effective.

3.2 Price Sensitivity Clustering

3.2.1 Data Preparation

For the Price Sensitivity Clustering the same pattern was used as in Load Pattern Clustering. The objective was to validate the presence of price sensitivity among users by examining whether the variance between electricity usage at High and Low tariff points at a given time could reveal distinct behavioral patterns. This approach assumes that households exhibiting greater variance in usage between the High and Low tariff times are more likely to be sensitive to price changes and opposite to that, those with less variance may be less responsive to such pricing dynamics. To this end, the variance for each observation point was calculated, creating a unique variance pattern, depicted in Figure 12, for each household.

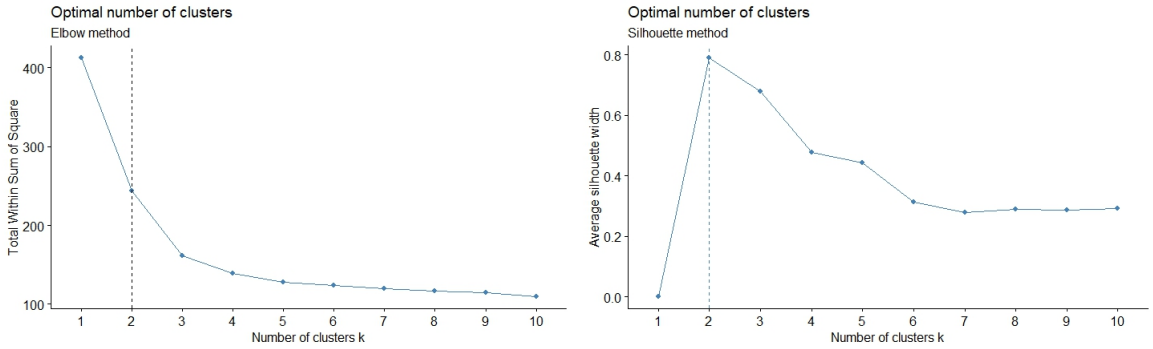

Figure 10: Cluster number validation

3.2.2 Clustering and Results

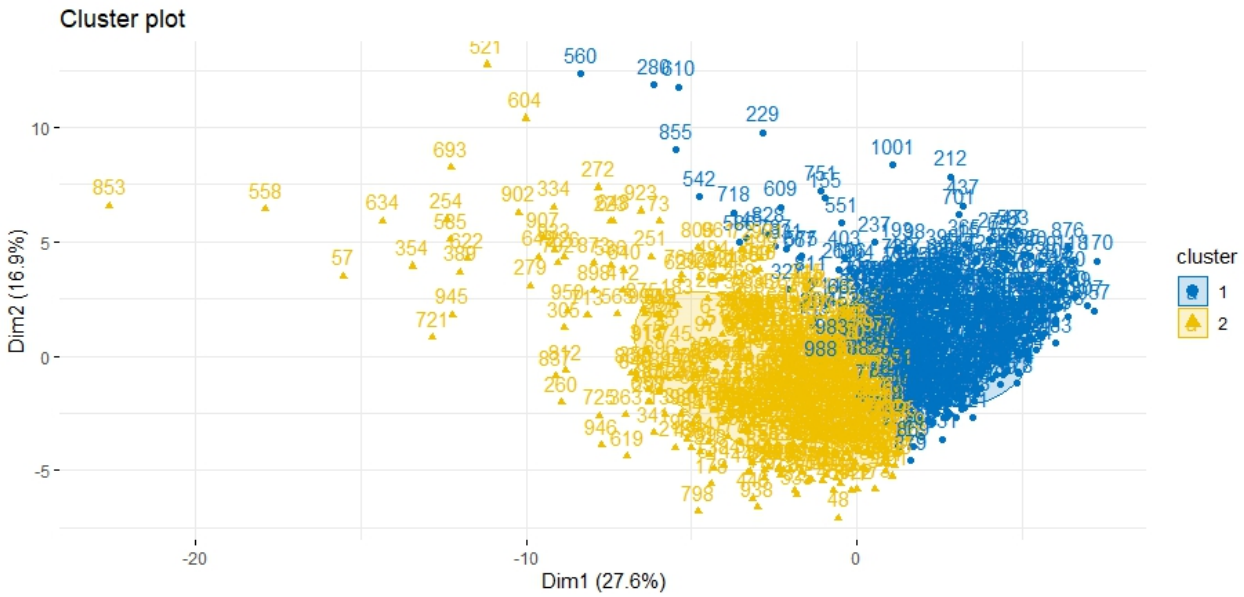

Figure 11: FCM cluster center plot.

To ascertain whether a household is price sensitive, a binary classification approach was adopted, simplifying the clustering into two distinct groups. Subsequent analyses using the Elbow and silhouette methods (Figure 10) confirmed the suitability of a two-cluster solution. From Fuzzy C-means cluster center plot two cluster separation can clearly be observed (Figure 11), however, higher degree of cluster overlapping is present. The cluster mean plot, as shown in

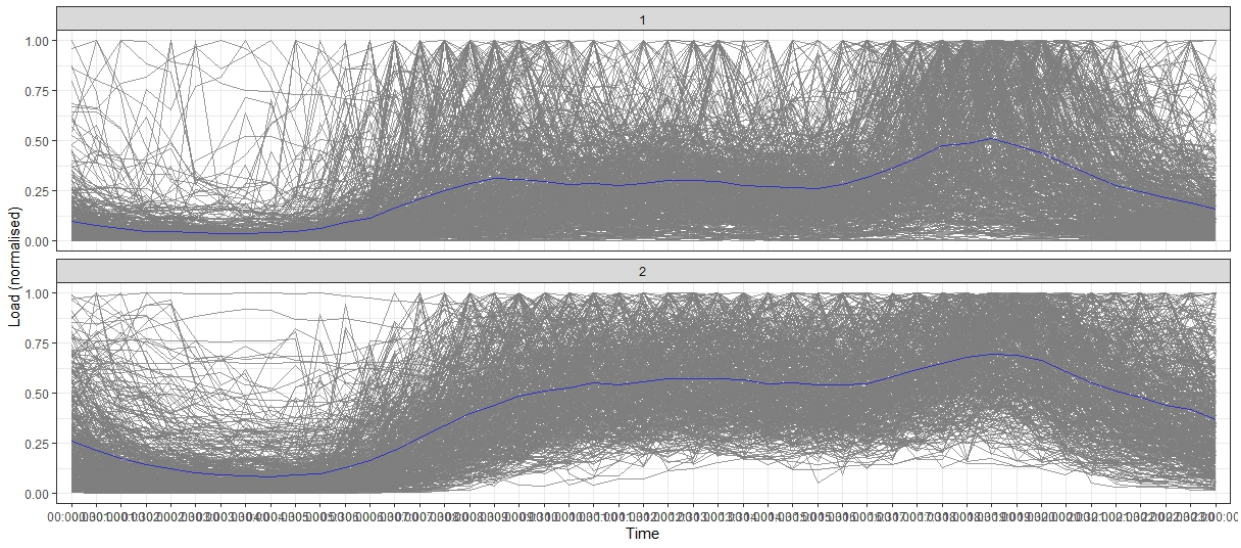

Figure 12: Clusters means.

Figure 12, offers further insight into the distinctiveness of two clusters identified by the clustering algorithm. Despite both clusters exhibiting similar variance patterns throughout the time, they

each possess distinguishable range differences. This observation aligns with the findings from the cluster center plot, reinforcing the conclusion that each cluster is distinguishable despite similarity in variance patterns across the clusters. This differentiation is crucial for further steps in identifying households that exhibit price-sensitive behaviors in response to price changes.

As before conjunction of CART model and Characteristics Index (Table 3) were used to classify households into clusters. Performance of classification was evaluated by comparing the predicted results from the classification tree to the actual clustering results for each household. Based on the results seen in Table 3 prediction accuracy was above 90% indicating that the CART model performed very effectively.

	1	2	%
1	467	42	91.7%
2	11	485	97.8%

Table 3: Classification Results

3.2.3 Result Validation

To validate whether the clustering results accurately identified households that were price-sensitive, a part of functional data analysis was conducted. To do so, once again normalized average daily consumption for two price tariffs for each household were calculated. To validated clustering results pointwise ANOVA analysis were conducted for both clusters separately with objective to investigated whether there were significant mean differences in the usage patterns between the two tariff rates within each cluster.

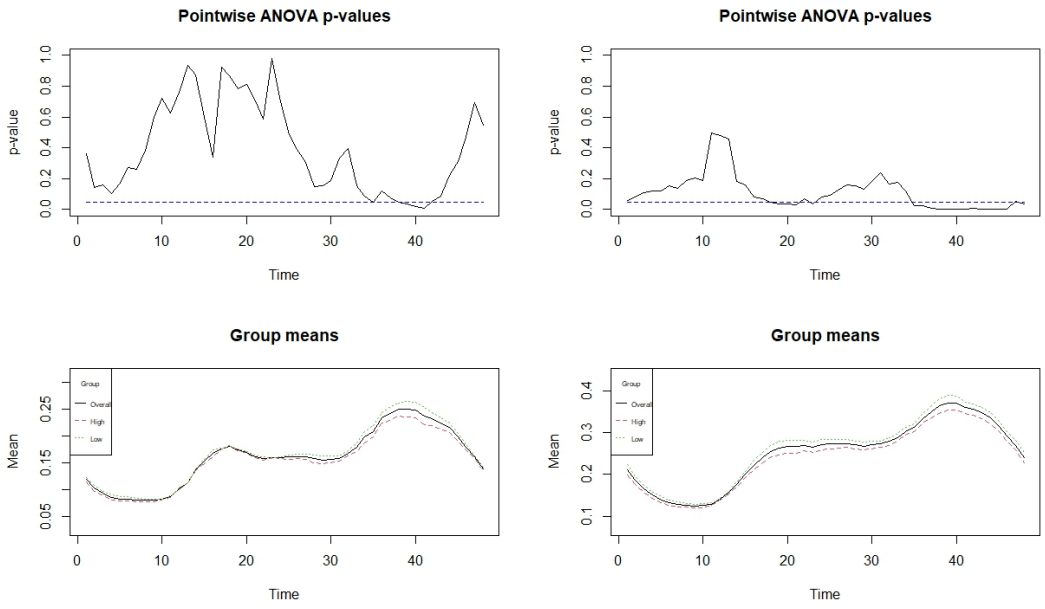

Figure 13: Clusters ANOVA analysis

The findings from the pointwise ANOVA analysis (Figure 13) revealed significant differences in variance in one of the clusters. This variance indicates that the households in this particular cluster exhibited notable differences in their energy consumption patterns when subjected to different tariff rates. This outcome suggests that these households are responsive to price changes, thereby validating the clustering approach.

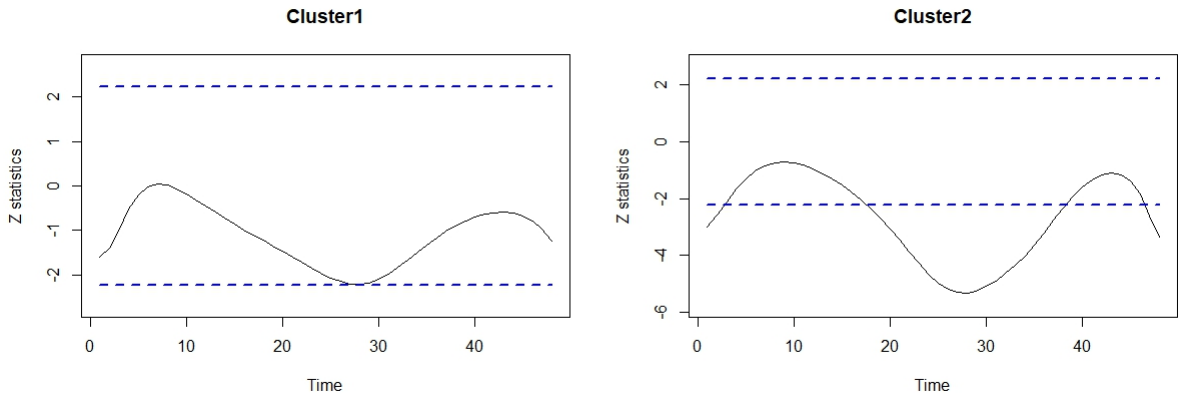

Figure 14: Two point test.

As in functional analysis to strengthen the findings from the pointwise ANOVA analysis, additional verification was carried out using two-sample pointwise tests. The results of these two-sample pointwise tests are presented in Figure 14. The testing process reaffirmed the initial conclusions drawn from the pointwise ANOVA analysis that significant variance observed in one of the clusters was not a result of random variation but a consistent pattern across the data-set.

4 Conclusions

4.1 Results

This paper presents an efficient two-stage customer segmentation methodology for electricity customer's classification based on load pattern recognition and customers' behaviour susceptibility to price changes.

The initial phase of the study validated the concept that users' electricity usage patterns can be indicative of price sensitivity. Concluded results provided interesting and somewhat unexpected results. Contrary to the logical presumption that the financial incentives would most significantly impact consumer behavior during peak energy consumption times, such as early morning or evening hours post-work, the significant deviations were determined at off peak hours. These observations indicate a variable price elasticity of electricity among consumers. The notable variations in consumption during off-peak times imply that customers' responsiveness to pricing is more multifaceted and less predictable than previously thought.

Building on these insights, the study introduces two-stage customer segmentation methodology. First stage provides and efficient way to classify customers into three different groups based on their load pattern and average electricity usage. Second stage introduced innovative way how to determine if customer's electricity sage is susceptibility to price changes, or in other words determine is price sensitive just from historical electricity usage data. This approach contrasts with numerous existing studies which have predominantly relied on qualitative data collected through surveys, or focused on macro-level patterns across the general population.

This paper shed light on the intricacies and dynamics of energy consumption behavior, underscoring the importance for energy providers and policymakers of considering the diverse and nuanced patterns of customer response to electricity pricing. This knowledge can be crucial when designing more effective and tailored pricing and policy strategies for either commercial interests and regulatory bodies.

4.2 Limitations and Further work

Looking ahead, the study opens possibilities for future research, particularly in exploring the broader implications of these findings on energy policy and market structuring. Further investigation could delve into how different demographic factors and regional characteristics or seasonality influence price elasticity and consumption patterns.

References

[1] ENA Lietuvos Energetikos Agentūra. 2023 m. lapkriČio mĖn. energetikos duomenŲ apŽvalga. https://www.ena.lt/uploads/Lauros/2023-11%20MAIN%20Energetikos%20duomenu%CC%A8%20apz%CC%8Cvalga_v1.pdf, December 2023.

[2] Jerzy Andruszkiewicz, Józef Lorenc, and Agnieszka Weychan. Demand price elasticity of residential electricity consumers with zonal tariff settlement based on their load profiles. *Energies*, 12(22), 2019.

[3] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems*, 18(1):381–387, 2003.

[4] L. G. Costacurta and M. A. Sanz-Bobi. Application of clustering methods for discovering patterns of energy use in regional areas for the residential sector. In *2017 IEEE Manchester PowerTech*, pages 1–6, 2017.

[5] Goutam Dutta and Krishnendranath Mitra. A literature review on dynamic pricing of electricity. *Journal of the Operational Research Society*, 68:1131–1145, 2017.

[6] V. Figueiredo, F. Rodrigues, Z. Vale, and J. B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems*, 20(2):596–602, 2005.

[7] Koichiro Ito. Do consumers respond to marginal or average price? evidence from nonlinear electricity pricing. *American Economic Review*, 104(2):537–63, February 2014.

[8] N. Bora Keskin, Yuexing Li, and Nur Sunar. Data-driven clustering and feature-based retail electricity pricing with smart meters. *SSRN Electronic Journal*, 2020.

[9] UK Power Networks. Smartmeter energy consumption data in london households. https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households, 2015.

[10] Y. Qi, B. Luo, X. Wang, and L. Wu. Load pattern recognition method based on fuzzy clustering and decision tree. In *2017 IEEE Conference on Energy Internet and Energy System Integration (EI2)*, pages 1–5, 2017.

[11] Benoît Ravel, Stéphane Auray, and Vincenzo Caponi. Price elasticity of electricity demand in france. *Economie et Statistique*, 513(1):91–103, 2019.

[12] ESO Energijos skirstymo operatorius. Per dieną daugiau nei 2 tūkst. išmaniųjų skaitiklių įdiegiantys eso inžinieriai: klientai jaučia skaitiklių naudą. `https://www.eso.lt/lt/ziniasklaida/per-diena-daugiau-nei-2-tukst.-ismaniuju-vw8f.html`, December 2023.

[13] Gianluca Trotta, Anders Rhiger Hansen, and Stephan Sommer. The price elasticity of residential district heating demand: New evidence from a dynamic panel approach. *Energy Economics*, 112:106163, 2022.

[14] G. J. Tsekouras, N. D. Hatziargyriou, and E. N. Dialynas. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Transactions on Power Systems*, 22(3):1120–1128, 2007.

[15] Y. Wang, L. Li, and Q. Yang. Application of clustering technique to electricity customer classification for load forecasting. In *2015 IEEE International Conference on Information and Automation*, pages 1425–1430, 2015.

5 Appendix A

```
#-------------------------------------------------------------------------------------
#libraries
library(fclust)
library(ppclust)
library(reshape2)
library(e1071)
library(imputeTS)
library(dplyr)
library(factoextra)
library(caret)
library(rpart)
library(ggplot2)
library(tidyr)
library(readr)
library(fda)
library(fds)
library(fda.usc)
library(fdANOVA)
library(mlr3misc)
#-------------------------------------------------------------------------------------
#Data Load
#xxxx need to change to folder path where files are stored
#-------------------------------------------------------------------------------------

base_path = "xxxx/LCL-June2015v2_"
file_number = 135:164

for (i in file_number) {
  file_name <- paste(base_path, i, ".csv", sep = "")
  var_name <- paste("suvart", i - (file_number[1]-1), sep = "")

  assign(var_name, read_csv(file_name,
                            col_types = cols(
                              DateTime = col_datetime(format = "%Y-%m-%d %H:%M:%S"),
                              `KWH/hh (per half hour)` = col_number()
                            )))
}

#Data aggregation to one data frame
#if using more than 30 files, please manually add data frame names to binding
#and removal scripts
suvart <-rbind(suvart1, suvart2, suvart3, suvart4, suvart5, suvart6, suvart7, suvart8,
               suvart9, suvart10, suvart11, suvart12, suvart13, suvart14, suvart15,
               suvart16, suvart17, suvart18, suvart19, suvart20, suvart21, suvart22,
```

```
               suvart23, suvart24, suvart25, suvart26, suvart27, suvart28, suvart29,
               suvart30)

#Remove separate frames
rm(base_path, file_name, file_number, i, var_name, suvart1, suvart2, suvart3, suvart4,
   suvart5, suvart6, suvart7, suvart8, suvart9, suvart10, suvart11, suvart12, suvart13,
   suvart14, suvart15, suvart16, suvart17, suvart18, suvart19, suvart20, suvart21,
   suvart22, suvart23, suvart24, suvart25, suvart26, suvart27, suvart28, suvart29, suvart30)


#Load Tariff Data
#xxxx need to change to folder path

tariff <- read_csv("xxxx/Tariffs.csv",
                   col_types = cols(TariffDateTime = col_datetime(format = "%m/%d/%Y %H:%M"),
                                    `Tariff` = col_character() ))

#Data rename and data type conversion
colnames(suvart)[1] <-"OBJ_ID"
colnames(suvart)[3] <-"DATE_ID"
colnames(suvart)[4] <-"CONSUMP"

colnames(tariff)[1] <-"DATE_ID"
colnames(tariff)[2] <-"PRICE"

suvart$TIME_ID <- format(suvart$DATE_ID,"%H:%M:%S")
suvart$DATE_ID <- as.Date(suvart$DATE_ID)

tariff$TIME_ID <- format(tariff$DATE_ID,"%H:%M:%S")
tariff$DATE_ID <- as.Date(tariff$DATE_ID)
tariff$stdorToU <- as.character("ToU")

#-------------------------------------------------------------------------------------
#Data prep for random cluster assignation
#Aggregating data to object(household) level and averaging data usage
#for observations at specific time
#-------------------------------------------------------------------------------------


Ccluster<-1:12

#Random cluster number assigned for each household
#Aggregation for each household group
nmlSUM = suvart %>%
  group_by(OBJ_ID, TIME_ID) %>%
```

```
  summarise_at("CONSUMP", list(mean))

dfX = dcast(nmlSUM, OBJ_ID ~ TIME_ID, value.var = "CONSUMP")

mapID = dfX[1]
mapID$Ccluster<-print(sample(Ccluster,nrow(mapID),replace=TRUE))

#Data set prepared with tariffs and random clusters
suvart2 <- left_join(suvart, tariff, by=c("stdorToU", "DATE_ID","TIME_ID"))
suvart2 <- left_join(suvart2, mapID, by=c("OBJ_ID"))
suvart2$PRICE = suvart2$PRICE %>% replace_na('None')

#Aggregation for each cluster + tariff group
nmlSUM = suvart2 %>%
  group_by(TIME_ID, PRICE, Ccluster) %>%
  summarise_at("CONSUMP", list(mean))
nmlSUM$PRICE = nmlSUM$PRICE %>% replace_na('None')
nmlSUM <- na.omit(nmlSUM)


dfX = dcast(nmlSUM, TIME_ID ~ Ccluster + PRICE, value.var = "CONSUMP")
dfX = na_replace(dfX, 0)

#-------------------------------------------------------------------------------------
#Price sensitivity analysis
#-------------------------------------------------------------------------------------

#Smoothing
monRng <- c(1,48)
bspl <- create.bspline.basis(monRng, norder=6)
Lfdobjelect = int2Lfd(6)
monfdPar <- fdPar(bspl, Lfdobjelect, 0.32)
plot(bspl, lwd=2)

monMatr <- as.matrix(dfX[,-1])
elect_fda <- smooth.basis(1:48, monMatr, monfdPar)

plot(elect_fda)
plot(monMatr)

plotfit.fd(monMatr, 1:48, elect_fda$fd)

#-------------------------------------------------------------------------------------
#mean Variation
vid <- mean.fd(elect_fda$fd)
```

```
std <- sd.fd(elect_fda$fd)
plot(elect_fda, lwd=1)
lines(vid, col = 2, lwd = 3)
lines(vid+std, col = "blue", lty=2, lwd=3)
lines(vid-std, col = "blue", lty=2, lwd=3)

#-------------------------------------------------------------------------------------
#Derivartives
#derivative 1
vel <- deriv.fd(elect_fda$fd, 1)
acc <- deriv.fd(elect_fda$fd, 2)
plot(vel, main = "First derivative")

#derivative 2
plot(acc, main = "Second derivative")

#-------------------------------------------------------------------------------------
#PCA
nharm = 2
elect_pca <- pca.fd(elect_fda$fd, nharm)
plot(elect_pca$harmonics, lwd=3)

plot.pca.fd(elect_pca, xlab='Time')

elect_V_pca <- varmx.pca.fd(elect_pca)
plot.pca.fd(elect_V_pca, xlab='Time')

#-------------------------------------------------------------------------------------
#Hypothesis testing
#-------------------------------------------------------------------------------------

#data frames split according to price groups
elect_fdaH <- Data2fd(1:48, monMatr[,c(1,5,9,13,17,21,25,29,33,37,41,43)],
                      basisobj=bspl)
elect_fdaL <- Data2fd(1:48, monMatr[,c(2,6,10,14,18,22,26,30,34,38,42)],
                      basisobj=bspl)
elect_fdaX <- Data2fd(1:48, monMatr[,c(3,7,11,15,18,23,27,31,35,39,43)],
                      basisobj=bspl)
elect_fdaN <- Data2fd(1:48, monMatr[,c(4,8,12,16,20,24,28,32,36,40,44)],
                      basisobj=bspl)

opar2 <- par(mfrow=c(2,2))
plot(elect_fdaH, xlab = "High")
plot(elect_fdaL, xlab = "Low")
plot(elect_fdaN, xlab = "Normal")
```

```
plot(elect_fdaX, xlab = "None")

#-------------------------------------------------------------------------------------
#fANOVA

dta.A <- elect_fdaH
t.sq <- seq(1,48, length=501)

#fAnova analysis
#assign groups
a <- c('High','Low','None','Normal')

gr <- factor(c(rep(a, 12)))
t.sq <- seq(1,48, length=48)

fANOVA.pointwise(data=monMatr, groups=gr, t.seq=t.sq, alpha=0.05)
plotFANOVA(x = monMatr, group.label = as.character(gr), int = c(0, 1), means = TRUE)

#-------------------------------------------------------------------------------------
#Two sample pointwise-test

opar2 <- par(mfrow=c(2,3))
stat <- Ztwosample(x=elect_fdaH, y=elect_fdaL, t.seq = t.sq, namesH = "High-Low")
stat <- Ztwosample(x=elect_fdaH, y=elect_fdaX, t.seq = t.sq, namesH = "High-None")
stat <- Ztwosample(x=elect_fdaH, y=elect_fdaN, t.seq = t.sq, namesH = "High-Normal")
stat <- Ztwosample(x=elect_fdaN, y=elect_fdaL, t.seq = t.sq, namesH = "Normal-Low")
stat <- Ztwosample(x=elect_fdaN, y=elect_fdaX, t.seq = t.sq, namesH = "Normal-None")
stat <- Ztwosample(x=elect_fdaL, y=elect_fdaX, t.seq = t.sq, namesH = "Low-None")


#-------------------------------------------------------------------------------------
#Load Pattern Clustering
#-------------------------------------------------------------------------------------

nmlSUM = suvart %>%
  group_by(OBJ_ID, TIME_ID) %>%
  summarise_at("CONSUMP", list(mean))
nmlSUM <- na.omit(nmlSUM)

dfX = dcast(nmlSUM, OBJ_ID ~ TIME_ID, value.var = "CONSUMP")
dfX = na_replace(dfX, 0)

#-------------------------------------------------------------------------------------
#checking for cluster center number
```

```
# Elbow method
fviz_nbclust(dfX[-1], kmeans, method = "wss") +
  #theme(text = element_text(size = 16)) +
  geom_vline(xintercept = 2, linetype = 2)+
  labs(subtitle = "Elbow method")

# Silhouette method
fviz_nbclust(dfX[-1], kmeans, method = "silhouette")+
  #theme(text = element_text(size = 16)) +
  labs(subtitle = "Silhouette method")

# Gap statistic - advised: very long run time
fviz_nbclust(dfX[-1], kmeans,  method = "gap_stat", nboot = 50)+
  #theme(text = element_text(size = 16)) +
  labs(subtitle = "Gap statistic method")

#-------------------------------------------------------------------------------------
#Clustering

sumData = dfX

#normalizing data
sumData[,-1] = t(apply(dfX[, -1], 1, function(x)(x)/(max(x))))
sumData = na.omit(sumData)

#Data split into training and test (sample_frac - proportion)
trData <- sumData %>% dplyr::sample_frac(1)
#tstData  <- dplyr::anti_join(sumData, trData, by = 'OBJ_ID')


#FCM, 3 clusters. 100-iteracions more info
#https://www.rdocumentation.org/packages/e1071/versions/1.7-3/topics/cmeans
xdf = cmeans(trData[,-1], 3, 100, FALSE,"euclidean", "cmeans", 2)


#Cluster center graph
fviz_cluster(list(data = trData[-1], cluster=xdf$cluster),
             ellipse.type = "norm",
             ellipse.level = 0.6,
             palette = "jco",
             ggtheme = theme_minimal())

trData$cluster = xdf$cluster
trData_long = gather(trData, time, measure, 2:49, factor_key=TRUE)
trData = trData[,-50]
```

```
#Cluster mean graph
ggplot(trData_long, aes(time, measure, group = OBJ_ID)) +
  facet_wrap(trData_long$cluster, ncol = 1, scales = "free_y") +
  geom_line(color = "grey10", alpha = 0.25) +
  geom_line(data = trData_long, aes(time, measure),
            color = "grey50", alpha = 0.60, size = 0.2) +
  stat_summary(fun.y=mean, group=1 , geom="line", colour="blue", alpha=0.6)  +
  labs(x = "Time", y = "Load (normalised)") +
  theme_bw()

#-------------------------------------------------------------------------------------
#Classification and regression tree
#Characteristics index
a1 <- (apply(trData[, -1], 1, function(x) mean(x)/(max(x))))
a2 <- (apply(trData[, -1], 1, function(x) min(x)/(max(x))))
a3 <- (apply(trData[, c(2,3,4,5,6,7,8,9,10,11,12,13)], 1,
             function(x) mean(x)/max(x)))
a4 <- (apply(trData[, c(14,15,16,17,18,19,36,37,38,39,40,41,42,43)], 1,
             function(x) mean(x)/max(x)))
a5 <- (apply(trData[, c(20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,44,45,46,47,48,49)],
             1, function(x) mean(x)/max(x)))


trData$a1 <- a1
trData$a2 <- a2
trData$a3 <- a3
trData$a4 <- a4
trData$a5 <- a5

rm(a1,a2,a3,a4,a5)
trData$cluster = xdf$cluster

opar2 <- par(mfrow=c(1,2))

#CART using characteristics index values
fit = rpart(xdf$cluster~., data = trData[,50:54], control = rpart.control(cp = 0.0001))
plot(fit)
text(fit, cex = 0.9, xpd = TRUE, digits = 1)
printcp(fit)

#Pruning
fit.pruned = prune(fit, cp = 0.0055)
plot(fit.pruned)
text(fit.pruned, cex = 0.9, xpd = TRUE, digits = 1)
```

```
#Testing
#Applying CART to same data-set
pred <- round(predict(fit.pruned, trData),0)

#Print matrix of clustering results vs CART
table(factor(trData$cluster, levels=min(pred):max(pred)),
      factor(pred, levels=min(pred):max(pred)))



#-------------------------------------------------------------------------------------
#Price Sensitivity Clustering and Result Analysis
#-------------------------------------------------------------------------------------

set.seed(25) #seed used to reproduce results

#removal of unnecesarry objects
rm(bspl, dta.A, elect_fda,elect_fdaH,elect_fdaL,elect_fdaN,elect_fdaX, Lfdobjelect,
   monfdPar,monMatr, stat, opar2, monRng, pred, t.sq, indexes, elect_pca, elect_V_pca,
   a, gr, nharm, dfX,dfX3, fit, fit.pruned,nmlSUM3, mapID, nmlSUM,nmlSUM4,std,sumData,
   trData,trData_long,vid,xdf, varin, vel, elect_fdaH1, elect_fdaL1, elect_fdaH2,
   elect_fdaL2, elect_fda1, elect_fda2, acc, grid_mat, mapID2, fns)

#random household clustering into one of 12 clusters
obj1 = suvart %>%
  group_by(OBJ_ID, TIME_ID) %>%
  summarise_at("CONSUMP", list(mean))

obj2 = dcast(obj1, OBJ_ID ~ TIME_ID , value.var = "CONSUMP")

mapID <- obj2[1]
mapID$Ccluster<-print(sample(Ccluster,nrow(mapID),replace=TRUE))
rm(obj1,obj2)

#final data set prepared
suvart3 <- left_join(suvart, tariff, by=c("stdorToU", "DATE_ID","TIME_ID"))
suvart3 <- na.omit(suvart3) #remove fixed rate observations

#unique(suvart3$PRICE)

suvart3$PRICE[suvart3$PRICE == 'Normal'] <- 'High' #combine Normal and High groups

#Aggregation for each cluster + tariff group
nmlSUM3 = suvart3 %>%
```

```
  group_by(OBJ_ID, TIME_ID) %>%
  summarise_at("CONSUMP", list(var)) #variance calculation

nmlSUM3 <- na.omit(nmlSUM3)

dfX3 = dcast(nmlSUM3, OBJ_ID ~ TIME_ID , value.var = "CONSUMP")
dfX3 <- na.omit(dfX3)

#-------------------------------------------------------------------------------------
#checking for cluster center number

# Elbow method
fviz_nbclust(dfX3[-1], kmeans, method = "wss") +
  #theme(text = element_text(size = 16)) +
  geom_vline(xintercept = 2, linetype = 2)+
  labs(subtitle = "Elbow method")

# Silhouette method
fviz_nbclust(dfX3[-1], kmeans, method = "silhouette")+
  #theme(text = element_text(size = 16)) +
  labs(subtitle = "Silhouette method")

# Gap statistic - advised: very long run time
fviz_nbclust(dfX3[-1], kmeans,  method = "gap_stat", nboot = 50)+
  #theme(text = element_text(size = 16)) +
  labs(subtitle = "Gap statistic method")

#-------------------------------------------------------------------------------------
#Clustering

#normalize data
sumData = dfX3
sumData[,-1] = t(apply(dfX3[, -1], 1, function(x)(x)/(max(x))))
sumData = na.omit(sumData)

trData <- sumData %>% dplyr::sample_frac(1.00)

#FCM, 2 cluster, 100 iterations
xdf = cmeans(trData[,-1], 2, 100, FALSE,"euclidean", "cmeans", 2)

#Cluster center graph
fviz_cluster(list(data = trData[-1], cluster=xdf$cluster),
             ellipse.type = "norm",
             ellipse.level = 0.6,
             palette = "jco",
```

```
            ggtheme = theme_minimal())

trData$Scluster = xdf$cluster
trData_long = gather(trData, time, measure, 2:48, factor_key=TRUE)
trData = trData[,-50]

#Cluster mean graph
ggplot(trData_long, aes(time, measure, group = OBJ_ID)) +
  facet_wrap(trData_long$Scluster, ncol = 1, scales = "free_y") +
  geom_line(color = "grey10", alpha = 0.25) +
  geom_line(data = trData_long, aes(time, measure),
            color = "grey50", alpha = 0.60, size = 0.2) +
  stat_summary(fun.y=mean, group=1 , geom="line", colour="blue", alpha=0.6)  +
  labs(x = "Time", y = "Load (normalised)") +
  theme_bw()

#-------------------------------------------------------------------------------------
#Classification and regression tree

#Characteristics index
a1 <- (apply(trData[, -1], 1, function(x) mean(x)/(max(x))))
a2 <- (apply(trData[, -1], 1, function(x) min(x)/(max(x))))
a3 <- (apply(trData[, c(2,3,4,5,6,7,8,9,10,11,12,13)], 1,
             function(x) mean(x)/max(x)))
a4 <- (apply(trData[, c(14,15,16,17,18,19,36,37,38,39,40,41,42,43)], 1,
             function(x) mean(x)/max(x)))
a5 <- (apply(trData[, c(20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,44,45,46,47,48,49)],
             1, function(x) mean(x)/max(x)))

trData$a1 <- a1
trData$a2 <- a2
trData$a3 <- a3
trData$a4 <- a4
trData$a5 <- a5

rm(a1,a2,a3,a4,a5)
trData$Scluster = xdf$cluster

opar2 <- par(mfrow=c(1,2))

#CART using characteristics index values
fit = rpart(xdf$cluster~., data = trData[,50:54], control = rpart.control(cp = 0.0001))
plot(fit)
text(fit, cex = 0.9, xpd = TRUE, digits = 1)
printcp(fit)
```

```
#Pruning
fit.pruned = prune(fit, cp = 0.01)
plot(fit.pruned)
text(fit.pruned, cex = 0.9, xpd = TRUE, digits = 1)

#Testing
#Applying CART to same data-set
pred <- round(predict(fit.pruned, trData),0)

#Print matrix of clustering results vs CART
table(factor(trData$Scluster, levels=min(pred):max(pred)),
      factor(pred, levels=min(pred):max(pred)))


#-------------------------------------------------------------------------------
#Validation of price sensitivity

mapID2 <- mapID

mapID <- left_join(trData[,c(1,55)], mapID, by=c("OBJ_ID"))

suvart3 <- suvart3[,1:6]
suvart3 <- left_join(suvart3, mapID, by=c("OBJ_ID"))

nmlSUM = suvart3 %>%
  group_by(TIME_ID, PRICE, Scluster, Ccluster) %>%
  summarise_at("CONSUMP", list(mean))
nmlSUM <- na.omit(nmlSUM)


dfX = dcast(nmlSUM, TIME_ID ~ PRICE+Scluster+Ccluster , value.var = "CONSUMP")
dfX = na_replace(dfX, 0)

opar2 <- par(mfrow=c(1,1))

#Smoothing
monRng <- c(1,48)
bspl <- create.bspline.basis(monRng, norder=6)
Lfdobjelect = int2Lfd(6)
monfdPar <- fdPar(bspl, Lfdobjelect, 0.32)
#plot(bspl, lwd=2)

monMatr <- as.matrix(dfX[,-1])
elect_fda <- smooth.basis(1:48, monMatr, monfdPar)
```

```
plot(elect_fda)
plot(monMatr)

plotfit.fd(monMatr, 1:48, elect_fda$fd)

#data split into price groups per cluster
#For PointwiseAnova
elect_fda1 <- monMatr[,c(1,2,3,4,5,6,7,8,9,10,11,12,25,26,27,28,29,30,31,32,33,34,35,36)]
elect_fda2 <- monMatr[,c(13,14,15,16,17,18,19,20,21,22,23,24,37,38,39,40,41,42,
                         43,44,45,46,47,48)]

#for Twosample test
elect_fdaH1 <- Data2fd(1:48, monMatr[,c(1,2,3,4,5,6,7,8,9,10,11,12)], basisobj=bspl)
elect_fdaL1 <- Data2fd(1:48, monMatr[,c(25,26,27,28,29,30,31,32,33,34,35,36)],
                       basisobj=bspl)

elect_fdaH2 <- Data2fd(1:48, monMatr[,c(13,14,15,16,17,18,19,20,21,22,23,24)],
                       basisobj=bspl)
elect_fdaL2 <- Data2fd(1:48, monMatr[,c(37,38,39,40,41,42,43,44,45,46,47,48)],
                       basisobj=bspl)

#-------------------------------------------------------------------------------
#mean Variation
vid <- mean.fd(elect_fda$fd)
std <- sd.fd(elect_fda$fd)
plot(elect_fda, lwd=1)
lines(vid, col = 2, lwd = 3)
lines(vid+std, col = "blue", lty=2, lwd=3)
lines(vid-std, col = "blue", lty=2, lwd=3)

#fANOVA
source("C:/Users/domas/Desktop/Magis/fANOVA.R")
t.sq <- seq(1,48, length=400)

#fAnova analysis
#assign groups
a <- c('High','Low')

gr <- factor(c(rep(a, 12)))
t.sq <- seq(1,48, length=48)

fANOVA.pointwise(data=elect_fda1, groups=gr, t.seq=t.sq, alpha=0.05)
fANOVA.pointwise(data=elect_fda2, groups=gr, t.seq=t.sq, alpha=0.05)
plotFANOVA(x = elect_fda1, group.label = as.character(gr), int = c(0, 1), means = TRUE)
```

```
plotFANOVA(x = elect_fda2, group.label = as.character(gr), int = c(0, 1), means = TRUE)

#Two sample pointwise-test

source("C:/Users/domas/Desktop/Magis/Ztwosample.R")
opar2 <- par(mfrow=c(1,2))
stat <- Ztwosample(x=elect_fdaH1, y=elect_fdaL1, t.seq = t.sq, namesH = "Cluster1")
stat <- Ztwosample(x=elect_fdaH2, y=elect_fdaL2, t.seq = t.sq, namesH = "Cluster2")



#-------------------------------------------------------------------------------------------
#Functions
#-------------------------------------------------------------------------------------------

fANOVA.pointwise <- function(data, groups, t.seq, alpha=0.05) {
  # data is matrix with time in rows and variables in columns
  # group is a list names separating columns into different groups, a factor
  # time scale for measures
  n <- nrow(data)
  pvals <- numeric(n)
  lv <- levels(groups)
  k <- length(lv)
  mean.p <- matrix(NA, ncol=k, nrow=n)
  perm <- factorial(k)/(factorial(2)*(factorial(k-2)))
  Tukey.posthoc <- matrix(NA, ncol=perm, nrow=n)
  for(i in 1:n) {
    dt <- data.frame((data[i,]), groups)
    names(dt) <- c("values", "groups")
    av <- aov(values~groups, data = dt)
    pvals[i] <- summary(av)[[1]]["Pr(>F)"][1,1]
    mean.p[i,]  <- as.matrix((dt %>% group_by(groups) %>% summarise(mean(values)))[,2])
    colnames(Tukey.posthoc) <- rownames(TukeyHSD(av)$groups)
    Tukey.posthoc[i,] <- TukeyHSD(av)$groups[,4]
  }
  overall_mean <- apply(data, 1, mean)

  opar1 <- par(mfrow=c(2,1))
  plot(t.seq, pvals, type="l", main = "Pointwise ANOVA p-values",
       xlab = "Time", ylab="p-value", ylim=c(0,1))
  lines(t.seq, rep(0.05, n), col="blue", lty=2)

  mn <- min(mean.p, overall_mean)
  mx <- max(mean.p, overall_mean)
```

```
  plot(t.seq, overall_mean, type = "l", main = "Group means",
       xlab = "Time", ylab = "Mean", ylim = c(mn-0.05, mx+0.05))
  for(i in 1:k) {
    lines(t.seq, mean.p[,i], col=i+1, lty=i+1)
  }
  legend("topleft", legend=c("Overall", lv), lty=1:(k+1), col=1:(k+1),
         cex=0.5,, title="Group")
  par(opar1)
  opar2 <- par(mfrow=c(2,3))

  for(i in 1:perm) {
    plot(t.seq, Tukey.posthoc[,i], type="l",
         main = paste("Tukey HSD p-values", rownames(TukeyHSD(av)$groups)[i]),
         xlab = "Time", ylab = "p-value", ylim = c(0,1))
    lines(t.seq, rep(0.05, n), col="blue", lty=2)
  }

  par(opar2)
  return(list(p.values=pvals, TukeyHSD=Tukey.posthoc, gr.means = mean.p,
              overal.mean=overall_mean))
}

#-------------------------------------------------------------------------------------------
# Two samples pointwise t-test

Ztwosample <- function(x, y, t.seq, alpha=0.05, namesH) {
  if(class(x) != "fd") stop("X must be fd object")
  if(class(y) != "fd") stop("Y must be fd object")
  k <- length(t.seq)

  mu.x <- mean.fd(x)
  mu.y <- mean.fd(y)

  n <- dim(x$coef)[2]
  m <- dim(x$coef)[2]

  delta <- (mu.x - mu.y)
  delta.t <- eval.fd(t.seq, delta)

  z.x <- center.fd(x)
  z.y <- center.fd(y)

  z.x.t <- eval.fd(t.seq, z.x)
  z.y.t <- eval.fd(t.seq, z.y)
  z.t <- cbind(z.x.t, z.y.t)
```

```
  if(n > k) {
    Sigma <- (t(z.t) %*% z.t)/(n-2)
  } else {
    Sigma <- (z.t %*% t(z.t))/(n-2)
  }

  gamma.t <- diag(Sigma)
  Zpointwise <- sqrt((n*m)/(n+m)) * delta.t/sqrt(gamma.t)

  crit <- qt(1-alpha/2, n-2)
  crit.val <- rep(crit, k)
  params <- list(critical.value = crit)


  mx <- max(cbind(Zpointwise, crit.val))
  mn <- min(cbind(Zpointwise, -crit.val))

  plot(t.seq, Zpointwise, type="l", xlab = 'Time', ylab = "Z statistics",
       main = namesH, ylim=c(mn-0.5, mx+0.5))
  lines(t.seq, crit.val, lty=2, lwd=2, col="blue")
  lines(t.seq, -crit.val, lty=2, lwd=2, col="blue")

  return(list(statistics.pointwise = Zpointwise,
              params = params))
}

#End
```