D2.1

R1.1: Dictionary of load profiles: methodology and results

LEGAL DISCLAIMER

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 891943.

DOCUMENT INFORMATION

Deliverable title	R1.1: Dictionary of load profiles: methodology and results
Dissemination level	Public
Submission deadline	31/12/2021
Version number	1
Authors	Carlos Quesada Granja (UD) Cruz Enrique Borges Hernandez (UD) Aritz Bilbao Jayo (UD) Chris Merveille (GOI) Leire Astigarraga (GOI) Noah Pflugradt (FZJ) Pablo Montero-Manso (University of Sydney, AB)
Internal reviewers	Panagiotis Fragkos (E3M) Thomas Nacht (4ER) Diego Casado Mansilla (UD)
External peer reviewers	Giacomo Marangoni (POLIMI) Irene Malvestio (EIEE) Maksymilian Kochański (PROAKADEMIA)
Document approval	Not needed
Scope of the document according to the DoA	The report will contain a description of the data sources, the data curation and anonymization process followed as well as a description of the segmentation procedure followed (using FFORMA methodology) and its results. Finally, the segmentation instrument and the resulting clusters will be included in this report.

EXECUTIVE SUMMARY

The heterogeneity of the energy sector at the residential level makes it difficult to develop specific strategies for fostering the energy transition towards renewable or more sustainable energy sources. Achieving a proper segmentation of residential consumers is of utmost relevance to determine the impact of potential policy instruments. This Deliverable presents a segmentation methodology, which creates a dictionary of electrical behaviour from the load profiles of households. This methodology can be roughly divided into four steps: (1) the processing of a wide sample of residential electrical load profiles in order to complete missing values, detect and correct inconsistencies, and provide a suitable format; (2) the extraction of features, understood as a *summary* of values of each load profile; (3) the application of several cluster analysis methods to the extracted features; and (4) the analysis of results to identify the most representative behaviour patterns as well as the assessment of regional and temporal trends among the elements. The household segmentation provided in this Deliverable will be used as both an external variable of the causal diagram in Task 2.2 and a key unit of assessment in Work Package 4.

The data sources collected consist of a pool of twenty datasets of load profiles from eight different countries, which have been obtained from online open access resources and non-publicly accessible electricity suppliers. The collected datasets are heterogeneous, i.e. with unequal lengths, sampling frequencies, and measurement units, and the way they store the load profile information varies from one to another. This implies performing different processing tasks per dataset. In addition, some datasets incorporate additional demographic and socioeconomic information that requires to be processed separately. The particular characteristics of each individual dataset have been summarised in a table indicating the number of supply points and their type, the country of origin, the sampling period, and the duration of the data collection process, among others. The five datasets selected for analysis include those labelled as (1) *Electric cooperatives* from Spain, (2) *ISSDA* from Ireland, (3) *Low Carbon London* from the UK, (4) *Elergone Energia* from Portugal, and (5) *NEEA* from the USA, totaling 33,114 supply points.

The *data processing* performed, that is all those actions for the extraction of relevant information from the original dataset files, including the tasks of *data extraction*, *data cleaning* and *feature extraction*, are then discussed. With regard to *data extraction*, the way in which the original dataset files are translated into *raw files* is explained. Raw files consist of useful information in a common and simple format. They only contain timestamped values of consumed electricity of individual households and are stored as comma-separated value (CSV) files. The *data extraction* process of the five selected datasets is thoroughly described. For example, the *Electric cooperatives* dataset has required dealing with the specificities of the Spanish electric market and its *Electrical Metering Information System* (SIMEL), while the complexity of the extraction of data from the rest of the collected datasets has been lower. Raw files, however, are still not ready for analysis, as these files may contain gaps of missing values, outliers, or other data inconsistencies that must be harmonised. This process is commonly known as *data cleaning*. Five operations have been mainly performed in this work: (1) data imputation, (2) adaptation to local time, (3) avoidance of COVID-19 lockdown dates, (4) exclusion of short load profiles, and (5) exclusion of 0-valued load profiles.

Data imputation is the process of replacing missing data. Two different strategies of data imputation are carried out depending on the length of the sequences of missing values. On the one hand, sequences of missing values accounting for eight consecutive hours or less of the time series are imputed by linear interpolation. On the other hand, sequences of missing values accounting for over eight consecutive hours of the time series are imputed by the Last Observation Carried Forward (LOCF) method using a 7-day season. This means that the missing samples are replaced with values from the preceding seven days, thus ensuring the preservation of the same time and day of the week. Regarding the adaptation to local time, in order to make the comparison between the different datasets possible, in this work all time series have been referenced to their local time zone, and the discontinuities produced by any daylight saving time have been eliminated. In addition, all time series running during COVID-19 lockdowns or *stay-at-home* periods have been split into two parts: a pre-COVID-19 one, excluding the lockdown period, and a post-COVID-19 one, including the in-lockdown and post-lockdown periods. Finally, all time series shorter than one year or with all their values equal to 0 have been excluded from the general processing.

Time series *feature extraction* is a dimensionality reduction technique that finds common characteristics in the data and provides a more manageable and representative subset of variables. It has a predominant role as a data processing tool. Essentially, feature extraction translates each load profile, regardless of its length or range of dates, into a reduced set of meaningful values, the so-called *features*, thus reducing the complexity of any subsequent processing. In total, 3,179 features are extracted from each time series, which can be categorised into seven types.

First are the *basic statistics*, which include statistical moments (mean, variance, skewness, kurtosis); quartiles and deciles; outlier-related statistics, such as the interquartile range; and the sum of all the values of the time series. Second, there are the so-called *seasonal aggregates*, which are the largest group of features. Features are obtained by splitting the time series into subsets and then calculating summary statistics for each. Fourteen groups of subsets are defined, each subset corresponding to particular time bands (e.g. hour intervals, days of the week, months, and more complex combinations of those). A subset comprises all samples of the time series for which their date and time fall within its time band. The summary statistics computed for each subset are the mean, the standard deviation, and the sum (in some groups). Third, the *peak and off-peak time bands*. These features are obtained by splitting the time series into subsets and then adding all samples of each subset together. Eight groups of subsets are defined, and the time band containing the maximum value is identified as the peak time band. Similarly, the time band with the minimum value is identified as the off-peak time band. Fourth, the *lag k-day autocorrelations*, that is the correlation between values that are k time periods apart. The selected values of k span from 1 to 28 days. Fifth are the load factors, typical from the electrical system analysis. The load factor is the average load divided by the peak load in a specified time period (days, weeks, and years). The last two groups are computed using predefined software packages: the *tsfeatures* R-package, which includes 64 features such as STL decompositions, autocorrelation coefficients, seasonal strengths, entropies, and other values resulting from different analyses; and the *catch22* package, which includes 22 features whose selection was based on their successful classification performance.

In addition to the regular features extracted directly from the time series, *metafeatures*, i.e. features extracted from supplementary information provided by the datasets, have also been included. These metafeatures are not necessarily numeric, as they incorporate information on the location of the recorded sites, the socioeconomic classification of the household, etc. Metafeatures are mainly used to label and describe certain characteristics of the time series and, therefore, of a particular cluster. Two groups of metafeatures can be distinguished: general and dataset-specific metafeatures. General metafeatures include the file names, the dataset IDs, the number of samples of the time series, etc. Data-specific metafeatures, however, depend on the metadata provided by each dataset. They include tariff types, socioeconomic classifications, survey responses, and so on.

Having described all the data processing tools (*data extraction*, *data cleaning*, and *feature extraction*), it is now time to detail the methods implemented. They are mainly three: (1) the methods used to perform an automatic clustering of the time series from a set of features; (2) the validation measures calculated to help select the number of clusters; and (3) a method for visualising the average content of each cluster.

After the reduction of all time series to a small set of features, *cluster analysis* is applied to automatically search for groups of related observations. The *clValid* R-package has been used for performing clustering. This package provides nine clustering algorithms, including hierarchical, partitional, neural networks-based, fuzzy, and model-based algorithms. In addition, *cluster validation measures* are provided. These measures are used to assess the quality of a given clustering analysis. Validation measures can be classified into two large groups: internal and stability measures. On the one hand, internal measures take only the dataset and the clustering partition as inputs, and use intrinsic information in the data to assess the quality of the clustering. The internal validation measures selected are the connectivity, the silhouette width, and the Dunn index. On the other hand, stability measures are a special version of internal measures. They evaluate the consistency of a clustering result by comparing it with the clusters obtained after each column is removed, one at a time. The stability validation measures selected are the average proportion of non-overlap, the average distance, the average distance between means, and the figure of merit.

Once the resulting clusters are known, the next step is to visualise their content in a way that facilitates the subsequent analyses. The content of a cluster can be represented as the average of all time series that compose that cluster. However, some difficulties arise when computing this average, since the time series involved are not of the same length and do not begin and end on the same dates. These requirements can be fulfilled by means of *heatmaps*. A heatmap is a 2D grid that shows with different colour intensities the variations in magnitude of a phenomenon of interest, such as, in this case, the electricity consumption in kWh. The horizontal axis of the heatmaps has been set to represent the days of the year, while the vertical axis represents the hours of the day. In this way, the average, as well as the standard deviation, of electricity consumption for a whole year can be observed at a glance.

Now that all the pieces of the puzzle have been briefly outlined, the strategy that has been chosen to obtain the most representative set of electricity consumption profiles is as follows: first, a series of cluster analysis has been performed on a selection of datasets, combining different clustering methods, sets of features, and cluster numbers. Then,

validation measures have also been calculated and represented. Finally, a hybrid 'metric driven–domain expert assessment' methodology has been used to select the most appropriate combination of time series, cluster analysis methods, sets of features, and number of clusters for obtaining the electricity consumption patterns.

Regarding the selection of time series, some common-sense criteria have been applied after the cleaning process, such as the exclusion of time series with a high percentage of imputed samples or those belonging to industrial or commercial sites. As for the selection of the most appropriate cluster analysis method, the nine proposed methods were run on several datasets separately to test their clustering efficiency. SOM arose as the algorithm providing more identifiable clusters in the eyes of the domain experts. Five sets of features were designed for the application of cluster analysis methods, including seasonal aggregates, electric tariffication seasonal aggregates, peak and off-peak periods, strengths and autocorrelations, and the *Catch-22* set. The seasonal aggregates provided the most identifiable cluster visualisations. Finally, the optimal number of clusters found depends on the dataset. In the case of all time series, 40 clusters have been used.

The analysis strategy has focused on obtaining a set of results that answer three fundamental questions: (1) which are the main patterns of electrical consumption common to all the datasets analysed; (2) which are the differences in electricity consumption between regions; (3) and, which are the differences in electricity consumption between the pre- and post-COVID-19 lockdowns stages. For the first question, most of the heatmaps obtained provide a very detailed picture of the electricity consumption of households, which in turn corresponds reliably to their activities and behaviour. Other heatmaps depict what appear to be offices, while others present sunlight-dependent patterns. In order to classify and categorise the heatmaps in an efficient way, a taxonomy has been developed to organise the power consumption patterns by type. The final taxonomy has been achieved by combining two approaches: one based on expert knowledge/citizen science activity, and another based on an automatically-generated hierarchy. Both approaches turned out to be unexpectedly similar. Alongside the taxonomy, the 40 patterns were labelled through expert knowledge. Lastly, a narrative description of the hypothetical household type behind each pattern has been developed.

For the second question (regional differences), the country composition of each cluster has been analysed, making it possible to perform a regional comparison. The assessment provided the following highlights: (1) two patterns associated with regular, all-day-at-home behaviour are among the most prevalent in all regions except Ireland; (2) the Irish dataset shows the greatest differences with respect to the rest; and (3), Portugal and the US share a pattern associated to SMEs that clearly dominates the dataset. As for the third question (time differences), 20 clusters were generated per period (pre- and post-COVID-19), and a matching strategy was followed to identify them with the 40 main patterns of electrical consumption. 17 out of 20 generated clusters with pre-COVID-19 load profiles have been found in the previously introduced 40 clusters, whereas 19 out of 20 have a match in the post-COVID-19 group. Certain effects of the strict mid-March 2020 lockdowns held in Spain, as well as the mobility restrictions of early-December, can be clearly observed in some of the heatmaps in the post-COVID-19 sub-dataset

Some ideas on future work and the potential exploitability of the developed methodology are explained at the end of this Deliverable.

TABLE OF CONTENTS

LIST OF ACRONYMS AND ABBREVIATIONS

Acronym	Long text
ADR	Average Day-Referred
ASCII	American Standard Code for Information Interchange
CBT	Customer Behaviour Trials
CNAE	National Classification of Economic Activities (from Spanish '*Clasificación Nacional de Actividades Económicas*')
CRU	Commission for Regulation of Utilities (Ireland)
CSV	Comma-Separated Values
CUPS	Unified Supply Point Code (from Spanish '*Código Unificado de Punto de Suministro*')
DHW	Domestic Hot Water
DNO	Distribution Network Operator
DST	Daylight Saving Time
HDF	Hierarchical Data Format
IQR	Interquartile Range
ISO	International Organisation for Standardisation
ISSDA	Irish Social Science Data Archive
LOCF	Last Observation Carried Forward
NA	Not Available
NACE	Statistical Classification of Economic Activities in the European Community
NEEA	Northwest Energy Efficiency Alliance
PCA	Principal Component Analysis
SIMEL	Electrical Metering Information System (from Spanish '*Sistema de Información de Medidas Eléctricas*')
SME	Small and Medium-sized Enterprises
TBR	Time Band-Referred
TS	Time Series
t-SNE	t-distributed Stochastic Neighbour Embedding
UTC	Coordinated Universal Time (from merging English and French abbreviations)
var	Volt-ampere reactive

1. Introduction

The energy transition requires the involvement of households and energy consumers in order to succeed. Fostering their participation is of paramount importance but current strategies tend to be a "one size fit all" solution that generally does not achieve the expected result as the residential sector is very heterogeneous. In fact, it is not even clear what is actually the best way to cluster it in order to create tailored solutions for each segment. Current Energy System Models (ESMs) aggregate all households of a country into one "representative agent", which is of course an oversimplification of the reality and may lead to incorrect, biassed analysis. In this context, it is clear that consumer segmentation is of vital importance as one solution is not going to fit all cases and different households will be impacted differently by policy instruments. In this Deliverable, we will present a methodology to segment households depending on their electrical load profile by creating a dictionary of electrical behaviour, inferring the main properties of each entry in the dictionary and, finally, assessing the regional and temporal distribution of the different elements. These load profiles can be then incorporated in the large-scale Energy System Models to increase their granularity and relevance for policy analysis.

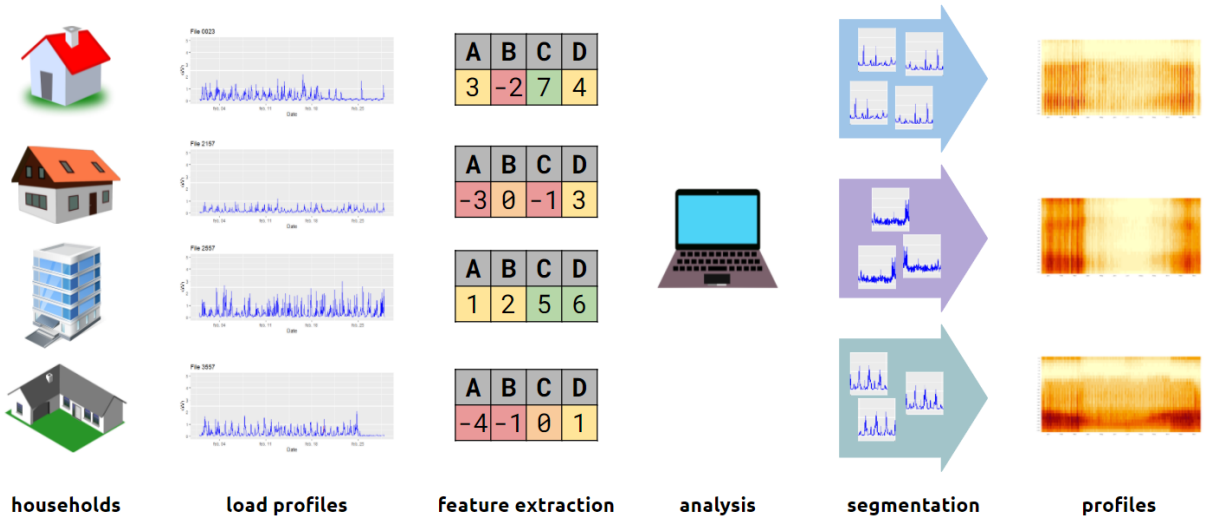


Figure 1: General overview of the segmentation process of household load profiles

Fig. 1 shows a graphic representation of the methodology followed, which is divided into a series of steps or procedures described below. The order of the sections in this document corresponds to those steps.

1. First, a broad sample of residential electrical load profiles are processed in order to complete missing values, detect data inconsistencies and outliers, and provide a proper format. Detailed description on the datasets analysed and the processing methods carried out can be found in Sections 2 and 3.
2. Then, a set of features is extracted from each time series. This set of features can be understood as a summary of the time series. Different subsets of features are defined in order to assess the most suitable one. A description of the features computed can be found in Section 4.

3. Then, different cluster analysis methods are applied to the features. Section 5 provides the details of the methods used and the methodology followed to optimise the hyperparameters of each of the cluster analysis methods.
4. Finally, Section 6 provides a description of the experiments carried out and the results obtained.

The results of this Deliverable will be the backbone of several activities in the WHY Project. In particular, the household segmentation will be used:

- as an external variable in the causal diagram on Task 2.2, in order to summarise the socioeconomic profiles;
- as fundamental units of assessment in WP4. In particular to assess what are the best forecasting methods for each segment (Task T4.1), optimise the set of technical and behavioural interventions for each consumer group (Task T4.2), and assess the impact on tariff changes (Task T4.3).

2. Data sources

A pool of twenty datasets of load profiles from seven European countries (and the US) has been gathered from online open access resources and non-publicly accessible electricity suppliers (provided by GOI and 4ER). Although the main focus is on the study of load profiles originating from households, the datasets collected are heterogeneous and contain, in addition to household data, load profiles originating from sub-metering equipment, aggregations of dwellings, blocks of buildings, and non-residential facilities, such as businesses, offices, companies, or public administration buildings.

Each dataset stores load profile information in its own way, using different data structures and file formats. Load profiles are stored as time series, which may have different lengths, sampling frequencies, and measurement units. This implies performing distinct processing tasks for each dataset. In addition, some datasets also incorporate additional information (metadata), such as demographic, sociological and economic information about the load profiles, which also need to be processed.

The time series contained in a dataset are usually circumscribed to a specific country and a specific data acquisition period, which normally encompasses several years. The number of sites or supply points recorded depends on the dataset, and can range from a few tens to several thousands. The particular characteristics of each individual dataset are summarised in Table 1.

Name	No. sites	Loc.	Site types	Sub-meter.	Sampling period	Collection period	TS length
Electric cooperatives	22 851	ES	**H**, C, I, F	no	1 h	2014-2021	2.5 yr*
EDRP	14 319	GB	H	no	30 min	2008-2010	2 yr*
SGSC	13 735	AUS	H	no	30 min	2012-2024	2 yr*
ISSDA	6 435	IE	**H**, C	no	30 min	2009-2010	1.5 yr*
SAVE	6 262	GB	H	no	15 min	2017-2018	2 yr*
Low Carbon London	5 449	GB	**H**, C, F	no	30 min	2011-2014	1.5 yr*
Kaggle	1 449	*v/l*	**B**, C, **F**	no	1 h	2016	1 yr
Elergone Energia	351	PT	H, C, I, F	no	15 min	2012-2014	2 yr*
NESEMP	215	GB	H	no	5 min	2010-2012	2 yr*
NEEA	200	US	H	yes	15 min	2020-2024	10 mo*
EnerNOC	100	US	**C**, I, F	no	15 min	2012	1 yr
Energy Retailer	100	AT	H	no	15 min	2020	10 mo
Technology Provider	99	AT	H	no	1 min	2019-2020	2 yr*
HTW Berlin	74	DE	H	no	1 s	*synthetic*	1 yr
ENLITEN	68	GB	H	yes	*variable*	2013-2016	1 yr*
GreenGrid	47	NZ	H	yes	*1s*	2016-2018	2 yr*
WPuQ	38	DE	H	no	1h	2018-2020	2 yr*
ADRES	30	AT	H	no	1 s	2011	14 d

LES	22	GB	H	no	1 min	2008-2009	2 yr*
REFIT	20	GB	H	yes	1 h	2013-2016	2 yr*
OPSD	11	DE	**H**, I, F	yes	15 min	2014-2019	1.5 yr*
Individual household	1	FR	H	no	1 min	2006-2010	47 mo

Table 1: Characteristics of the datasets gathered.

The meaning of the columns in Table 1 are indicated below:

- ***Name***: name used to identify the dataset.
- ***No. sites***: number of data collection sites or supply points. All submetering measures belonging to the same household are counted as a single collection site.
- ***Loc.***: country of the collection sites by their ISO 2-letter code. *v/l* stands for "various locations".
- ***Site types***: H (households), C (commercial sites, including businesses and offices), I (industrial sites), F (educational, healthcare, cultural, or governmental facilities), and B (buildings or household aggregates). Bold text indicates the site type with the highest number of time series within the dataset
- ***Submeter.***: indicates whether the dataset includes utility submetering time series.
- ***Sampling period***: sampling period of the time series. *Variable* means that sampling does not follow a fixed period for the same time series.
- ***Collection period***: interval of years in which the data were collected. *Synthetic* means that the time series have been artificially created by processing other pre-existing time series and are, therefore, not linked to any real collection period.
- ***TS length***: duration of the time series. Asterisks (*) indicate that the dataset contains time series of different lengths, the value being an estimate of the median duration.

Although the ultimate goal is to analyse all datasets in Table 1, priority has been given to datasets with the highest number of time series, and coming mostly from households. The geographical diversity of the selection has also been sought.

Therefore, this has excluded from preliminary analyses datasets with a low number of sites, or those originating from entire buildings or utility submetering. For obvious reasons, all datasets collected at later dates of performed analysis (e.g. EDRP, SAVE, NESEMP, WPuQ, LES) have also been excluded from the original selection of datasets. Their descriptions, though, have been added to Table 1 for the sake of completeness.

The five datasets selected for analysis include (1) *Electric cooperatives*, (2) *ISSDA*, (3) *Low Carbon London*, (4) *Elergone Energia*, and (5) *NEEA*, totaling 33 114 different sites or supply points.

3. Data processing

This section describes all actions carried out for the extraction of relevant information from the original dataset files, which contain the household load profiles, and their subsequent transformations prior to analysis. From a technical point of view, the complete process by which information is extracted from the original dataset files is summarised in the following diagram:

Figure 2: Information extraction process from original datasets files.

Data processing involves the first four steps of the sequence, which are described throughout this Section and the following. First, it is necessary to inspect the dataset files to know how data is structured. Each dataset is different in terms of file types (including plain text files, Excel spreadsheets, CSV files, HDF files, etc.), data formats, and the units of the stored values (kWh, kW, var, etc.). Therefore, it is not possible to implement a single solution to extract the relevant information from all datasets at a time. It is necessary to create a script (or several of them) for each dataset to *translate* the useful information into a common and simple format, which is referred to here as *raw* files. This process is known as **data extraction** and is analysed in Section 3.1. Next, raw files usually present missing values or unordered samples that need to be corrected and harmonised. This process is called **data cleaning**, and is detailed in Section 3.2. Finally, prior to cluster analysis and visualisation, all processed time series are subjected to **feature extraction**. Due to the particular relevance of this technique, a separate section (Section 4), is devoted to it.

3.1. Data extraction

This section explains how original dataset files are translated into **raw files**. These files only contain timestamped values of consumed electricity of individual households. Raw files are stored as CSV files, i.e. plain text files with two comma-separated fields: the timestamps in the format `YYYY-MM-DD hh:mm:ss`, and the value of electricity consumption in kWh. If data on electric self-production is available, a third field next to electricity consumption also appears (in kWh).

All the scripts used for data extraction can be found in the **`whyT2.1`** package[1], at the GitHub repository of the main contributor partner of this task. The R programming language[2] has been mainly used for the creation of scripts, although Python[3] has also been used. The data extraction process of the five selected datasets is described below.

[1] https://github.com/DeustoTech/why-T2.1
[2] https://www.r-project.org/about.html
[3] *Van Rossum, G., & Drake Jr, F. L. (1995). Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam.*

3.1.1. Spanish electric cooperatives

This is one of the most complex datasets processed in terms of extraction of relevant information. The dataset is made up of 35 862 files (nearly 37 GB of data) following the format described by the Spanish '*Electrical Metering Information System*' (SIMEL)[4].

In total, up to 6 years (at best) of hourly data have been provided from 22,851 supply points, ranging from 2013 to 2021. Their locations are concentrated in the regions of Basque Country and Navarre (around 90% of the dataset), although representation from all regions of peninsular Spain exists, encompassing both metropolitan and rural areas. In addition to the location of the customers, the dataset provides information on the contracted tariff and the '*National Classification of Economic Activities*' (CNAE) code[5], which nominally identifies the households. Although not exactly the same, CNAE codification has some equivalence to the NACE codification ('Statistical Classification of Economic Activities in the European Community')[6] used in the EU.

In order not to make the reading of this document more complicated, all cumbersome technical procedures regarding data extraction from the original SIMEL files can be found in **Annex A**.

3.1.2. Irish Social Science Data Archive (ISSDA)

The *Irish Social Science Data Archive* (ISSDA)[7] is the leading centre for the acquisition, preservation and dissemination of quantitative data in Ireland. It is hosted by the University College Dublin Library and its main objectives are to ensure access to quantitative datasets in the social sciences, and to promote international comparative studies of the Irish economy and society.

Among the datasets currently held by ISSDA, the study known as '*0012-00 Commission for Energy Regulation*'[8] contains relevant information for the WHY project, especially the part related to the "*Electricity Customer Behaviour Trial*". As stated on the ISSDA website, during 2009 and 2010 some '*Smart Metering Electricity*' customer behaviour trials were carried out, involving more than five thousand Irish households and businesses. The aim of the trials was to assess the impact on consumers' electricity consumption, in order to inform the cost-benefit analysis for a national rollout. Electric Ireland's residential and business customers who participated in the trials were provided with a smart electricity metre in their homes (or facilities), and agreed to participate in the research to establish how smart metering can help shape energy usage behaviours across a variety of socio-economic variables. The trials produced positive results, which are available from the *Commission for Regulation of Utilities* (CRU), along with more information on the '*Smart Metering Project*'.

[4] https://www.ree.es/sites/default/files/01_ACTIVIDADES/Documentos/Documentacion-Simel/SIMEL_Ficheros_Intercambio_Informacion_v37.pdf
[5] https://www.ine.es/daco/daco42/clasificaciones/cnae09/cnae_2009_rd.pdf
[6] https://www.ine.es/daco/daco42/clasificaciones/rev.1/correspondencia_nace.pdf
[7] https://www.ucd.ie/issda/
[8] *Commission for Energy Regulation (CER). (2012). CER Smart Metering Project - Electricity Customer Behaviour Trial, 2009-2010. 1st Edition. Irish Social Science Data Archive. SN: 0012-00.* https://www.ucd.ie/issda/data/commissionforenergyregulationcer/

The detailed data underlying the electricity customer behaviour trial results is available in anonymised format to facilitate further research. A '*Data Request Form for Research Purposes*' has to be completed and signed in order to access the data.

Details on data extraction from the original dataset files can be found in **Annex A**.

3.1.3. Low Carbon London

'*Low Carbon London*'[9] was the name of a 4-year innovation project to investigate the impact of a wide range of low-carbon technologies on London's electricity distribution network. The project was led by the *UK Power Networks* DNO between November 2011 and February 2014. Energy consumption readings for a sample of 5,567 London households took part in the project. Readings were taken at half hourly intervals. The customers in the trial were recruited as a balanced sample representative of the Greater London population. **Annex A** contains information on data extraction from the original dataset files.

3.1.4. Elergone Energia

The 'Electricity Load Diagrams 20112014' dataset available at the *UCI Machine Learning Repository*[10] contains the electricity consumption of 370 supply points (*clients*, as specified in the repository) of the Portuguese retailer 'Elergone Energia'. Each time series is composed of up to four years of complete quarter-hourly records (2011-2014). No metadata files are provided, so the dataset does not indicate whether the readings come from households or have other origins such as businesses or industrial sites. The only extra data provided is basic information to perform the data extraction process correctly. The consumption of each customer, in kW, appears as a separate column in a large CSV file, where each row is a timestamp. Time series starting later than January 1, 2011 are padded with zeros.

3.1.5. Northwest Energy Efficiency Alliance (NEEA)

As part of the *End Use Load Research* (EULR) project[11], the NEEA manages a dataset of residential load profiles, available upon request, covering the Northwestern United States. The project, currently ongoing (as of Sept. 2021), runs from 2020 to 2024, covers 200 homes (out of 400 planned), and provides quarter-hourly data. The dataset also includes disaggregated power data for an extensive list of appliances within each house. Among the metadata files available in the dataset, information can be found on location, time zones, and thorough descriptions of the measured appliances.

[9] https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households
[10] https://archive.ics.uci.edu/ml/datasets/ElectricityLoadDiagrams20112014
[11] https://neea.org/data/nw-end-use-load-research-project/energy-metering-study-data

3.2. Data cleaning

At this point, all load profiles have been extracted from all datasets of interest and saved as CSV files. These load profiles are referred to as *raw* files, indicating that the only processing performed on them has been their extraction as-is from the datasets.

However, these raw files are still not ready for analysis. These files may contain gaps of missing values, data inconsistencies, or outliers that must be corrected and harmonised. This process is usually known as data cleaning. The transformation from raw files to processed files is discussed in this Section. Five operations are mainly performed:

1. data imputation,
2. adaptation to local time,
3. avoidance of COVID-19 lockdown dates,
4. exclusion of short load profiles, and
5. exclusion of 0-valued load profiles.

These operations are explained in the following sections.

3.2.1. Data imputation

Missing values can appear in two different ways in the CSV raw files: (1) explicitly, as missing values of consumed energy for particular timestamps; or (2) implicitly, as missing tuples timestamp-value. In both cases, missing values are expressed as NA (not available) values when read by any data-centric programming language, such as R[12]. Two different strategies of data imputation are carried out depending on the length of the sequences of missing values.

On the one hand, NA sequences accounting for eight consecutive hours or less of the time series are imputed by linear interpolation. On the other hand, NA sequences accounting for over eight consecutive hours of the time series are imputed by the Last Observation Carried Forward (LOCF) method using a 7-day season. This means that the missing samples are replaced with values from the preceding seven days, thus ensuring the preservation of the same time and day of the week. If NA sequences are longer than seven days, the same sequence is repeated as much as required. In case missing samples are located at the beginning of the time series, where no previous seven days exist, they are replaced with values obtained from the next non-NA period of seven days.

Imputation algorithms mentioned here are already implemented in the `imputeTS` R package[13]. For illustrative purposes, Table 2 summarises the number of raw files that successfully completed all five operations in the data cleaning process ("Processed TS" column), their average length ("TS length" column), and the percentage of imputed samples per dataset ("Imputation of TS" column).

[12] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[13] https://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf

Name	Processed TS (No.)	TS length (in days)	Imputation of TS (%)
Electric cooperatives (pre-COVID-19 dataset)	12 123	921 ± 114	0.3 ± 3.6
ISSDA	6 088	801 ± 0	0.2 ± 2.4
Low Carbon London	5 269	801 ± 3	0.2 ± 0.9
Elergone Energia	351	1 238 ± 215	0
NEEA	64	801 ± 0	1.2 ± 2.0
Total	23 895	868 ± 113	0.2 ± 2.9

Table 2: Characteristics of the time series processed.

3.2.2. Adaptation to local time

Datasets in which timestamps are linked to Daylight Saving Time (DST) show one-hour discontinuities in spring and autumn, thus significantly altering most of the electricity consumption patterns. This is especially noticeable in all daylight-dependent load profiles. Whereas some datasets are referenced to coordinated universal time (UTC), which is not altered by DST, others are referenced to their local time zone, which generally follows some form of DST. In addition, datasets expressed in UTC usually show a time offset corresponding to their local time zone.

To make the comparison between the different datasets possible, in this work all datasets have been referenced to their local time zone but eliminating the discontinuities produced by any DST.

3.2.3. Avoidance of COVID-19 lockdown dates

The lockdowns adopted at the beginning of 2020 in several countries as an exceptional COVID-19 containment measure forced a drastic change in the household energy consumption habits. In this work, datasets whose time series run during COVID-19 lockdowns have been splitted into two sub-datasets: a **pre-COVID-19 dataset** that excludes the lockdown period, and a **post-COVID-19 dataset** that includes the in-lockdown and post-lockdown periods.

As shown in Table 1, this only occurs in the Spanish electric cooperatives and NEEA datasets. For the former, the cut-off date has been set at March 1, 2020, as lockdowns in Spain were implemented on March 15. For the latter, the cut-off date has been set at March 15, 2020, as different implementations of the *stay-at-home* orders in the US states of the dataset started around March 24, 2020. Because of the short length of the time series (see next section), no post-COVID-19 sub-dataset has been extracted from the NEEA dataset.

3.2.4. Exclusion of short load profiles

The minimum length of the time series to be processed is set to one year. This is explained by the fact that, during the processing, some of the features to be extracted require all time series to have at least annual length. Although time series shorter than one year could have been subjected to an extension process similar to that of the imputation of long sequences of missing values, it has been preferred to keep the time series unchanged. Time series shorter than one year have been excluded from the general data processing.

3.2.5. Exclusion of 0-valued load profiles

Although extremely rare, some time series have been found where all their values are 0. Since this makes some features impossible to compute, all 0-valued time series have been excluded from the general processing.

4. Feature extraction

As shown in Fig. 2, **feature extraction** is part of the data processing of the time series but, because of its particular relevance to this work, it has been separated out as a dedicated section.

Overall, the pool of load profiles obtained from the datasets can be viewed as a set of thousands of time series, with different start and end dates, that makes it complex to perform homogeneous analyses. Time series feature extraction is a **dimensionality reduction technique** that finds common characteristics in the data and provides a more manageable and representative subset of variables. Essentially, feature extraction translates each load profile, regardless of its length or range of dates, into a reduced set of meaningful values, the so-called features, thus reducing the complexity of any subsequent processing.

4.1. Procedure

Feature extraction is conducted in this work using the `tsfeatures`[14] and `catch22`[15] R-packages, which provides methods to extract features from time series data. To harmonise the time series of the different datasets, all of them have been pre-processed to consist of one sample per hour. This implies aggregating contiguous values when the sampling period is shorter than an hour, as in the case of datasets with half-hourly (e.g. ISSDA, Low Carbon London), or quarter-hourly (e.g. Elergone Energia, NEEA) time series.

The computed features can be roughly divided into two categories: those that are dependent on the amplitude of the time series, and those that are not (see Table 3). When the features to be computed are not amplitude-dependent, the time series are scaled to mean 0 and standard deviation 1 prior to feature extraction. Otherwise, time series are left as is. It is worth noting that the amplitude of the time series analysed here refers to the energy consumed in kWh by a household.

4.2. Extracted features

In total, 3 179 features can be extracted from each time series. Features can be categorised into seven types, which are summarised in Table 3 and explained in detail in the following sections.

Feature category	Standardisation?	No. of features
Basic statistics	no	20
Seasonal aggregates	no	3 014
Peak time bands	no	25
Lag *k*-day autocorrelations	yes	28

[14] *Hyndman R, Kang Y, Montero-Manso P, Talagala T, Wang E, Yang Y, O'Hara-Wild M (2022). tsfeatures: Time Series Feature Extraction. R package version 1.0.2.9000, https://pkg.robjhyndman.com/tsfeatures/*
[15] *Trent Henderson (2021). catch22: Calculation of 22 CAnonical Time-series CHaracteristics. R package version 0.1.3. https://github.com/benfulcher/catch22*

Load factors	no	6
`tsfeatures` R-package	yes	64
`catch22` R-package	yes	22

Table 3: Summary of feature types.

The features described below have been calculated for all time series of the selected datasets and included in the **feature dataset**. This dataset can be found in the **Zenodo repository** with the title of '*Time Series from Smart Meters*'.[16]

4.2.1. Basic statistics

Features that can be considered as basic statistics are shown below. These features are amplitude-dependent and, therefore, the time series have not been standardised for their calculation.

- **Statistical moments**. Quantitative measures related to the shape of the probability distribution of the time series. This includes the mean, variance, skewness, and kurtosis of the time series. The names given to the features are `mean`, `variance`, `skewness`, and `kurtosis`, respectively.
- **Quartiles**. Each of the five values which divide the population of values of the time series into four equal groups. This includes the minimum and maximum values of the time series (Q_0 and Q_4, respectively), the median (Q_2), and the lower and upper quartiles (Q_1 and Q_3, respectively) of the time series. The names given to the features are `minimum`, `maximum`, `median`, `quartile_1`, and `quartile_3`, respectively.
- **Deciles**. Each of the eleven values which divide the population of values of the time series into ten equal groups. This includes deciles D_1 to D_4, and D_6 to D_9. Deciles D_0, D_5 and D_{10} are not stored because they directly correspond to Q_0, Q_2 and Q_4, respectively. The names given to the features are `decile_x`, with `x` being an integer number ranging from 1 to 9, excluding 5.
- **Outlier-related statistics**. This includes the interquartile range (IQR), which is a measure of the statistical dispersion computed as $IQR = Q_3 - Q_1$; and the percentage of outliers of the time series, obtained from the IQR criterion for outlier detection[17]. The names given to the features are `iqr` and `iqr_outlier_pc`, respectively.
- The **sum** of all the values of the time series. For sums by specific time bands, see "*Seasonal aggregates*". The name given to the feature is `sum`.

4.2.2. Seasonal aggregates

This section describes the bulk of the features extracted. Features are obtained by splitting the time series into subsets and then calculating summary statistics for each. Fourteen groups of subsets are defined, each subset corresponding to particular time bands (see Table 4). A subset comprises all samples of the time series for which their date and time fall within its time band.

[16] https://zenodo.org/record/4455198
[17] Tukey, John W. Exploratory data analysis. Vol. 2. 1977.

For instance, the subset "3 pm" in Group I includes all samples of the time series for which their time is 15:00, regardless of the date (recall that, at this point, all time series are sampled on an hourly basis). Likewise, the subset "4 am to 7 am in summer" in Group VIII includes all samples of the time series for which their time is 4:00, 5:00, 6:00 or 7:00 and date is between June 1 and August 31 (meteorological summer).

Group	Time bands	Subsets
I	**1-hour intervals**: "12 am", "1 am", …, "10 am", "11 am", "12 pm", "1 pm", …, "10 pm", and "11 pm".	24
II	**Consecutive 4-hour intervals from midnight onwards**: "12 am to 3 am", "4 am to 7 am", "8 am to 11 am", "12 pm to 3 pm", "4 pm to 7 pm", and "8 pm to 11 pm".	6
III	**Consecutive 6-hour intervals from midnight onwards**: "12 am to 5 am", "6 am to 11 am", "12 pm to 5 pm", and "6 pm to 11 pm".	4
IV	**Days of the week**: "Sunday", "Monday", "Tuesday", …, and "Saturday".	7
V	**Weekday or weekend**: "weekday" and "weekend".	2
VI	**Months**: "January", "February", …, "November", and "December".	12
VII	**Northern hemisphere meteorological seasons**: "spring", "summer", "autumn", and "winter".	4
VIII	**Consecutive 4-hour intervals from midnight onwards per Northern hemisphere meteorological season**: "12 am to 3 am in spring", "4 am to 7 am in spring", "8 am to 11 am in spring", …, and "8 pm to 11 pm in winter".	24
IX	**Time bands of the Spanish 2.0TD electricity tariff excluding weekends**[18]: "12 am to 7 am", "8 am to 9 am", "10 am to 1 pm", "2 pm to 5 pm", "6 pm to 9 pm", and "10 pm to 11 pm".	6
X	**Time bands of the Spanish 2.0TD electricity tariff**: "12 am to 7 am", "8 am to 9 am", "10 am to 1 pm", "2 pm to 5 pm", "6 pm to 9 pm", "10 pm to 11 pm", and "weekend".	7
XI	**Time bands of the Spanish 2.0TD electricity tariff excluding weekends per Northern hemisphere meteorological season**: "12 am to 7 am in spring", "8 am to 9 am in spring", "10 am to 1 pm in spring", …, and "10 pm to 11 pm in winter".	24
XII	**Time bands of the Spanish 2.0TD electricity tariff per Northern hemisphere meteorological season**: "12 am to 7 am in spring", "8 am to 9 am in spring", "10 am to 1 pm in spring", "2 pm to 5 pm in spring", …, and "weekend in winter".	28
XIII	**Periods of the Spanish 2.0TD electricity tariff**: "*punta*" (peak), "*valle*" (standard), and "*llano*" (off-peak).	3
XIV	**Periods of the Spanish 2.0TD electricity tariff per month per year from 2014 to 2022**: "*punta* in January 2014", "*valle* in January 2014", "*llano* in January 2014", "*punta* in February 2014", …, "*llano* in December 2022".	108

Table 4: The fourteen groups of subsets defined for the 'seasonal aggregates' features.

[18] https://datadis.es/en/tarifas

The summary statistics computed for each subset are the mean and the standard deviation in all Groups, and also the sum in Groups IX to XIV. All features are provided both as is (absolute values), and as percentages of their Group (relative values). For consistency, relative standard deviations are computed referring to their corresponding values of relative means.

The features in Groups V to VIII, which are made up of time bands longer than a day, are computed in two different ways:

1. the usual one, by computing the summary statistics of all subsets; and
2. by subdividing each subset into one-day blocks and calculating the summary statistics of all the blocks of each subset.

While, in the first case, features are obtained for the entire time band (time band-referred values, or **TBR values**), in the second case, values are obtained for an average day of the time band (average day-referred values, or **ADR values**).

For compatibility with previous feature versions, TBR features are in turn divided by the number of days of their corresponding time bands (in Groups including February, it is considered as having 28.25 days). These features are called **TBR features per day**, and are a rough approximation of the ADR features. Features in Groups X to XIV, which are also made up of time bands longer than a day, are only computed as ADR values. No TBR values or TBR values per day are computed for them.

Annex B contains the naming conventions for all seasonal aggregates features.

4.2.3. Peak and off-peak time bands

Features are obtained by splitting the time series into subsets and then adding all samples of each subset together. Eight groups of subsets are defined, each subset corresponding to the time bands defined in Groups I to VIII of Table 4. As in the previous section, a subset comprises all samples of the time series for which their date and time fall within its time band.

For each Group, the time band containing the maximum value is identified as the **peak time band**, and the time band containing the minimum value is identified as the **off-peak time band**. Unlike *seasonal aggregates*, where the values of the features are not necessarily integers (as they represent means, variances and sums), *peak and off-peak time bands* are integer labels corresponding to a time band. The chosen integer to represent a time band is representative: e.g. 19 for the subset "7 pm" in Group I; 2 for the subset "Monday" in Group IV; 10 for the subset "October" in Group VI, and so on (see column '*Time bands and representative integer*' in Table 5). As in the previous section, TBR features per day are also computed for groups V to VIII.

The names given to the features are summarised in the following table. TBR features add the suffix _pday to the feature name:

Group	Time bands and representative integer	Feature names

I	**1-hour intervals**: "12 am" = 0, "1 am" = 1, …, "10 am" = 10, "11 am" = 11, "12 pm" = 12, "1 pm" = 13, …, "10 pm" = 22, and "11 pm" = 23.	`peak_hour_1`, `off_peak_hour_1`
II	**Consecutive 4-hour intervals from midnight onwards**: "12 am to 3 am" = 0, "4 am to 7 am" = 4, "8 am to 11 am" = 8, "12 pm to 3 pm" = 12, "4 pm to 7 pm" = 16, and "8 pm to 11 pm" = 20.	`peak_hour_4`, `off_peak_hour_4`
III	**Consecutive 6-hour intervals from midnight onwards**: "12 am to 5 am" = 0, "6 am to 11 am" = 6, "12 pm to 5 pm" = 12, and "6 pm to 11 pm" = 18.	`peak_hour_6`, `off_peak_hour_6`
IV	**Days of the week**: "Sunday" = 1, "Monday" = 2, "Tuesday" = 3, …, and "Saturday" = 7.	`peak_day`, `off_peak_day`
V	**Weekday or weekend**: "weekday" = 2, and "weekend" = 7.	`peak_weeday`, `off_peak_weekday`
VI	**Months**: "January" = 1, "February" = 2, …, "November" = 11, and "December" = 12.	`peak_month`, `off_peak_month`
VII	**Northern hemisphere meteorological seasons**: "spring" = 3, "summer" = 6, "autumn" = 9, and "winter" = 12.	`peak_season`, `off_peak_season`
VIII	**Consecutive 4-hour intervals from midnight onwards per Northern hemisphere meteorological season**: "12 am to 3 am in spring" = 300, "4 am to 7 am in spring" = 304, "8 am to 11 am in spring" = 308, …, and "8 pm to 11 pm in winter" = 1220.	`peak_hour4_season`, `off_peak_hour4_season`

Table 5: The eight groups of subsets defined for the 'peak timebands' features.

4.2.4. Lag k-day autocorrelations

The correlation coefficient between two values in a time series is known as the autocorrelation function. It is a way to measure the linear relationship between an observation at time *t* and the observations at previous times. A lag *k* autocorrelation is the correlation between values that are *k* time periods apart. In this case, the period considered is the day, i.e. 24 hourly samples. The correlation between values of the time series that are *k* days apart is computed. The selected values of *k* span from 1 to 28 days. The names given to the features are `ac_day_k`, with `k` being an integer number ranging from 1 to 28.

4.2.5. Load factors

In the electrical system analysis, the load factor is the average load divided by the peak load in a specified time period. Features have been computed for selected time periods of days, weeks and years. For each time period, the mean and standard deviation of all computed load factors in the time series are provided. The names given to the features are `load_factor_XX_YY`, with `XX` being mean for the mean, and `sd` for the standard deviation; and `YY` being `day`, `week`, and `month` depending on the specified time period.

4.2.6. tsfeatures package

All predefined methods for feature extraction provided by the function `tsfeatures()` in its June 2020 version are applied to the time series, with the exception of those methods that either take excessive computation time (in the order of tens of seconds per time series), or require input parameters, or return values that are not appropriate as features. Computed features include STL decompositions, autocorrelation coefficients, seasonal strengths, entropies, and other values resulting from different analyses. See the documentation of this R-package for a description of the key features.

4.2.7. Catch22

Catch22 is a high-performing subset of 22 features of the over 7 000 features in the *HCTSA* software package. Features were selected based on their classification performance across a collection of 93 real-world time-series classification problems, as described in a successful paper[19]. The twenty-two *Catch22* features are included in the extracted features.

4.3. Metadata features (metafeatures)

In addition to the regular features extracted directly from the time series, other features extracted from the metadata, which can be called **metafeatures**, have been included. These metafeatures are not necessarily numeric, as they incorporate information on the location of the recorded sites, the socioeconomic classification of the household, and so on. Metafeatures are mainly used to label and describe certain characteristics of the time series and, therefore, of a particular cluster.

Two groups of metafeatures can be distinguished: general and dataset-specific metafeatures. **General metafeatures** include:

- `file`: a string indicating the name of the file (without extension) where the time series is stored.
- `data_set`: a 3-letter string encoding the originating dataset name of the time series, e.g. “goi” for the Spanish electric cooperatives datasets, “iss” for the ISSDA dataset, “lcl” for the Low Carbon London dataset, “por” for the Elergone Energia dataset, “nee” for the NEEA dataset, and so on.
- `num_of_samples`: an integer value indicating the total number of samples of the imputed time series, i.e. that with all its NA values replaced.
- `ts_start_date`: a string indicating the date, time, and timezone of the first sample of the time series. The format is “`YYYY-MM-DD hh:mm:ss UTC`”.
- `ts_end_date`: a string indicating the date, time, and timezone of the last sample of the time series. The format is the same as in the previous field.
- `ts_days`: a real number indicating the length of the time series in days.
- `abs_imputed_na`: an integer indicating the number of imputed samples of the time series.

[19] Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., & Jones, N. S. (2019). catch22: Canonical time-series characteristics. *Data Mining and Knowledge Discovery*, 33(6), 1821-1852.

- `rel_imputed_na`: a real number indicating the percentage of imputed samples with respect to the total number of samples of the imputed time series.
- `mdata_file_idx`: an integer indicating the index (row number) of the metadata file from which the time series metadata has been extracted.
- `country`: a 2-letter string encoding the country of the time series (ISO 3166[20]).
- `is_household`: a binary value expressed by integers 0 (FALSE) and 1 (TRUE) indicating whether the time series belongs to a household. This differentiation is possible in those datasets whose metadata distinguish the origin of the time series. If not, this value is NA.

Data-specific metafeatures depend on the metadata provided by each dataset, and the names given to them are self-explanatory. For example, some of the metafeatures found in the Spanish electric cooperatives dataset are the municipality, the CNAE code, and the tariff type, to mention a few. In the *Low Carbon London* dataset, the ACORN socioeconomic types and groups have been added. Some data-specific metafeatures appear in more than one dataset at a time, e.g. the "`administrative_division`". For the ISSDA dataset, many of the answers to the questions from the pre-trial survey (recall Section 3.1.2) have been included as metafeatures. The list of included questions is comprehensive, and can be consulted in **Annex C**.

[20] https://www.iso.org/iso-3166-country-codes.html

5. Methods applied

This section focuses on the cluster analysis performed and describes:

1. the methods used to perform an automatic clustering of the time series from a set of features;
2. the validation measures calculated to help select the number of clusters; and
3. the method of visualising the average content of each cluster.

5.1. Cluster analysis methods

After the time series have been reduced to a small set of features (their number will depend on the selected set of features, between 5 and 25, see section 6.1.3), cluster analysis is applied to automatically search for groups of related observations.

The `clValid` R package[21] has been used for performing clustering. This package provides nine clustering algorithms, including:

- **hierarchical algorithms**[22], such as UPGMA (Unweighted Pair Group Method with Arithmetic mean), and DIANA (DIvisive ANAlysis);
- **partitional algorithms**, such as *k*-means[23], PAM (Partitioning Around Medoids)[24], and CLARA (Clustering LARge Applications);
- algorithms based on **unsupervised neural networks**, such as SOM (Self-Organising Maps)[25], and SOTA (Self-Organising Tree Algorithm)[26];
- the FANNY **fuzzy algorithm**;
- and the **model-based clustering**[27], based on parameterized finite Gaussian mixture models.

The `clValid::clValid()` function computes the different clustering algorithms by calling functions already implemented in other packages (except for SOTA). The calls to the functions that compute the *k*-means and FANNY algorithms within the `clValid` package have been modified to accommodate the characteristics of the inputs. These modifications can be include, for the `kmeans()` function:

- the maximum number of iterations allowed (`iter.max`) has been increased from 10 to 200;
- the number of random sets (`nstart`) has been increased from 1 to 50;
- and the computing `algorithm` has been changed from `Hartigan-Wong` to `MacQueen`.

[21] *Brock, G., Pihur, V., Datta, S., & Datta, S. (2008). clValid: An R package for cluster validation. Journal of Statistical Software, 25, 1-22.*
[22] *L. Kaufman and P. J. Rousseeuw. Finding Groups in Data. An Introduction to Cluster Analysis. Wiley, New York, 1990*
[23] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).
[24] *Van der Laan, M., Pollard, K., & Bryan, J. (2003). A new partitioning around medoids algorithm. Journal of Statistical Computation and Simulation, 73(8), 575-584.*
[25] *T. Kohonen. Self-Organizing Maps. Springer-Verlag, second edition, 1997.*
[26] *J. Dopazo and J. M. Carazo. Phylogenetic reconstruction using a growing neural network that adopts the topology of a phylogenetic tree. Journal of Molecular Evolution, pages 226–233, 1997.*
[27] *C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association, 17:126–136, 2001.*

Likewise, for the `cluster::fanny()` function:

- the maximum number of iterations (`maxit`) has been increased from 500 to 2 000;
- and the *membership exponent* (`memb.exp`) has been reduced, as suggested by some warning messages, from 2 to 1.05.

The calls to the rest of the functions have been kept with their default parameters. These modifications can be found in the **`clValid2`** package[28], at the GitHub repository of the main contributor of this task.

5.2. Cluster validation measures

The `clValid` R package also includes cluster validation measures. These measures are used to assess the quality of a given clustering analysis of one (or some) dataset(s). Validation measures can be classified into two large groups: internal and stability measures:

- **Internal** measures take only the dataset and the clustering partition as input and use intrinsic information in the data to assess the quality of the clustering.
- **Stability** measures are a special version of internal measures. They evaluate the consistency of a clustering result by comparing it with the clusters obtained after each column is removed, one at a time.

For internal validation, the selected measures are:

- **Connectivity**: can be roughly described as the number of nearest neighbours that belong to a different cluster. It has a value between zero and infinite and should be minimised.
- **Silhouette width**: for each observation, the silhouette width is the difference between the distance to the closest observation in another cluster and the average distance to the elements of the same cluster that the observation belongs to. A good clustering solution has large values for the average silhouette width for each observation in the dataset. Its solution lies in the interval [–1, 1], and should be maximised.
- **Dunn index**: the smallest distance among observations in different clusters divided by the largest distance among observations in the same cluster. The larger the ratio the more stable the clustering. It has a value between zero and infinite, and should be maximised.

Stability validation measures are a special version of internal measures which evaluate the clustering result by comparing it with the clusters obtained by removing a column (i.e. a feature) at a time. For stability validation, the selected measures are:

- **Average proportion of non-overlap (APN)**: measures the mean proportion of observations not placed in the same cluster. Its value lies in the interval [0, 1], with values close to zero corresponding with highly consistent clustering results.

[28] https://github.com/quesadagranja/clValid2

- **Average distance (AD)**: measures the mean distance among observations in the same cluster. It has a value between zero and infinite, and smaller values are preferred.
- **Average distance between means (ADM)**: measures the mean distance between cluster centres for observations placed in the same cluster. Its values range between zero and infinite, and again smaller values are preferred.
- **Figure of merit (FOM)**: measures the intra-cluster variance of the removed column. Its value ranges between zero and infinite, with smaller values equaling better performance.

The `clValid` R package includes a third type of validation measures called "**biological measures**", which is more suitable for biological (genetic) clustering tasks. They have not been used in this project.

5.3. Cluster visualisation

Once the resulting clusters are known, the next step is to visualise their content in a way that facilitates the subsequent analyses. The clustered elements are nothing more than points, i.e. numeric values (features) extracted from the time series, in a multidimensional space. Plotting these points as 2D or 3D projections (using PCA, t-SNE, or any other dimensionality reduction method that facilitates high-dimensional data visualisation) would not give a general idea of the type of time series that features represent. Rather than the values of the features themselves, it is **the average time series** behind each cluster that is of interest to be displayed.

The content of a cluster can be represented as the average of all time series that compose that cluster. However, some difficulties arise when computing this average, since the time series involved are not of the same length and do not begin and end on the same dates. In addition, the average should cover the longest seasonal period (one year) while preserving the structure of the shorter seasonal periods of the time series (days and weeks). This implies not to overlap different times or different days of the week.

One of the ways to fulfil these requirements is by means of **heatmaps**. A heatmap is a 2D grid that shows with different colour intensities the variations in magnitude of a phenomenon of interest, such as, in this case, the electricity consumption in kWh. The horizontal axis of the heatmaps has been set to represent the days of the year, while the vertical axis represents the hours of the day. In this way, the average electricity consumption for a whole year can be observed at a glance.

The computation of the average time series of a cluster is based on the week date representation described in the international standard ISO 8601[29]. This standard sets the possibility to express any particular date as an unambiguous triplet of integers, consisting of

1. the **ISO year**, which is slightly offset to the Gregorian year and, for certain dates, they do not match;

[29] https://www.iso.org/iso-8601-date-and-time-format.html

2. the **week number**, between 1 and 52 or 53 (depending on the year), which results in years with 364 or 371 days instead of the usual 365 or 366 days; and
3. the **weekday number**, between 1 and 7, with 1 being Monday, and 7 being Sunday.

By way of example, Tuesday, December 31, 2019 has the triplet

- ISO year = 2020,
- week number = 1,
- weekday number = 2,

whereas Thursday, December 31, 2020 (one year later) has the triplet

- ISO year = 2020,
- week number = 53,
- weekday number = 4.

Note they belong to the same ISO year but have extreme week numbers. This example is a special case as years with 53 ISO weeks (or long years) are uncommon. Among the years included in the collected datasets, only **2009**, **2015** and **2020** are long years.

Considering the maximum year length possible (371 days of 53-week years), the hourly samples of any year of any time series can be rearranged into a 24 x 371 matrix, with rows representing the hours of the day (from 00h to 23h), and columns representing the dates of that ISO year (see Fig. 5). The column $c(D)$ in which to arrange the hours of a particular date D can be calculated as

$$c(D) = 7 \cdot (w(D) - 1) + d(D),$$

where $w(D)$ is the week number of the date D, and $d(D)$ is the weekday number of the date D.

***w*(*D*)**	1							2							···	53						
***d*(*D*)**	1	2	3	4	5	6	7	1	2	3	4	5	6	7	···	1	2	3	4	5	6	7
wday	Mo	Tu	We	Th	Fr	Sa	Su	Mo	Tu	We	Th	Fr	Sa	Su	···	Mo	Tu	We	Th	Fr	Sa	Su
***c*(*D*)**	1	2	3	4	5	6	7	8	9	10	11	12	13	14	···	365	366	367	368	369	370	371
00h															···							
01h															···							
02h															···							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋮	⋮
23h															···							

Figure 5: Heatmap scheme: the 24 x 371 matrix in which the samples of a time series are arranged.

This ensures that samples with the same day and month, but from different years, are placed in columns as near to each other as possible while **strictly respecting** the day of the week. This is of utmost importance for our analyses, as it highlights the differences in electricity consumption between weekdays and weekends. Indeed, columns 1, 8, 15, …, ($7n$ + 1), …, 365 always correspond to Mondays; columns 2, 9, 16, …, ($7n$ + 2), …, 366 to

Tuesdays; and so on. The labels of the month names displayed on the horizontal axis of the resulting heatmaps (check Fig. 6) are, therefore, merely indicative.

The procedure by which **a single time series** of any length is converted into a single matrix is as follows:

1. The time series is standardised (mean 0 and standard deviation 1) and split into blocks according to its ISO year. Blocks with incomplete years are allowed.
2. Each block is rearranged as a 24 x 371 matrix. Empty matrix elements, i.e. containing no samples, (if any) are filled in with NA values.
3. The mean of all matrices is calculated on an element-by-element basis, ignoring all NA values (if any).

Due to the fact that most of the years are 52-week years, it was decided to complete their *fictitious* 53rd week with the first week of the following ISO year (if possible). This prevents the last week of most time series from being blank.

The procedure by which **a set of single matrices** (each representing a single time series, e.g. those contained in the same cluster) is converted into a single visualisation matrix is as follows:

1. Each matrix is standardised (mean 0 and standard deviation 1).
2. The mean of all matrices is calculated on an element-by-element basis, ignoring NA values if any.

This final visualisation matrix represents the **mean** of all time series in a cluster, and can be displayed using any function that creates a grid of colored pixels from its values. R function `image()` has been used here. The **standard deviation** of all time series in a cluster has been also computed using a procedure similar to the one described above. Colours ranging from yellow to red have been used for displaying the mean matrix, whereas different shades of blue have been used for the standard deviation matrix (see Fig. 6).

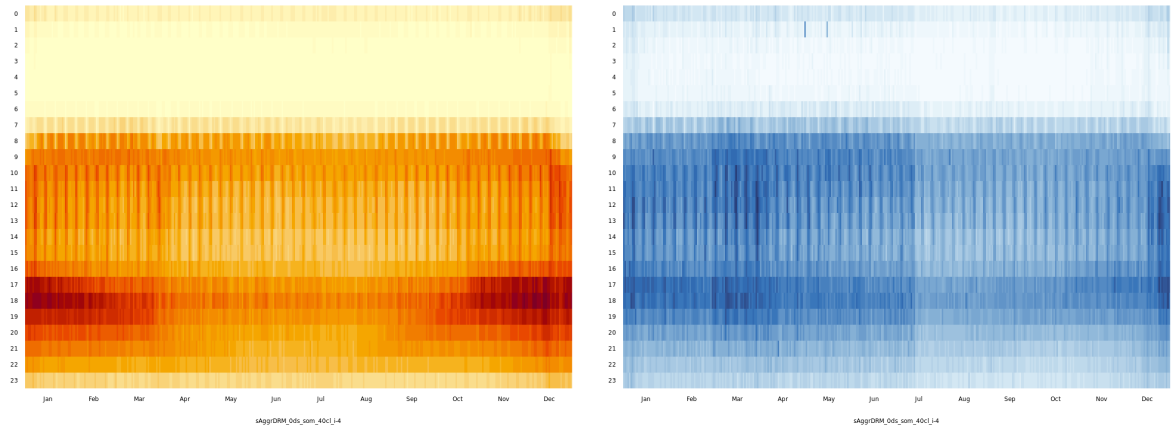


Figure 6: Mean (left) and standard deviation (right) visualisation matrices (or heatmaps) of a particular cluster.

It is worth noting that one of the side-effects of aligning dates according to their day of the week is that the electricity consumption on public holidays, which is usually similar to that of weekends, tends to appear scattered over a few days in a heatmap of the mean. This is because public holidays generally occur on the same day of the year, but the day of the week varies over the years calculated.

6. Results and discussion

The strategy chosen to obtain the most representative set of electricity consumption profiles is as follows: first, a series of cluster analyses has been performed on a selection of datasets, combining all clustering methods described in Section 5.1, different sets of features, and for various numbers of clusters. In addition, validation measures described in Section 5.2 have also been calculated and represented. Then, the expert analysis, supported by the validation measures, have led to the selection of a set of profiles that represent very well all possible residential electricity consumption patterns across European countries.

Different types of graphs have been created for the evaluation of the clusters obtained:

- Heatmaps, as described in Section 5.3, for the visualisation of an average year of energy consumption in a cluster.
- Box-plots for displaying the feature values in a cluster. The box-plots are represented with whiskers within the 1.5 IQR value.
- Bar-plots for the metafeature values in a cluster.

In addition, line charts and dendrograms have been used to display the validation measures and similarities of each set of clusters.

6.1. Methodology selection

This section discusses the selection of the most appropriate combination of datasets, cluster analysis methods, sets of features, and number of clusters for obtaining the electricity consumption patterns. For this purpose, a hybrid 'metric driven–domain expert assessment' methodology has been used. On the one hand, objective quality metrics for each one of the decisions have been computed, and the most promising solutions were presented to a panel of domain experts for their assessment. The panel of expert[30] was composed on:

- **Carlos Quesada Granja (UD)**: Machine learning expert and main developer of the activity.
- **Cruz Enrique Borges Hernandez (UD)**: Researcher with more than 10 years of experience on forecasting residential and commercial load profiles.
- **Chris Merveille (GOI)**: Engineer and physicist, senior researcher, coordinator of the GoiEner R&D projects team.
- **Leire Astigarraga (GOI)**: Engineer in renewable energies and expert in electricity market analysis.
- **Noah Pflugradt (FZJ)**: Main developer of a tool for the simulation of realistic residential behaviour patterns.
- **Pablo Montero-Manso (University of Sydney, Advisory Board)**: Expert on statistical methods for validation of models.

[30] *Contact emails, in order of appearance: carlos.quesada@deusto.es, cruz.borges@deusto.es, chris.merveille@goiener.com, leire.astigarraga@goiener.com, n.pflugradt@fz-juelich.de, and pablo.monteromanso@sydney.edu.au.*

6.1.1. Selection of time series

All time series of the selected datasets (see Section 2) have been subjected to the cleaning process described in Section 3.2, resulting in 23 895 time series of at least one year in length, with no missing samples, referenced to their local time zones, and occurring prior to the COVID-19 lockdowns. The 9 219 discarded time series did not contain more than one year of data. According to the literature reviewed, such a large number of different electrical load profiles have never been gathered, processed, or analysed before.

For the analyses described in this section, the database of 23 895 time series has been further refined on the basis of the following criteria:

- First, time series classified as non-household, i.e. those `is_household` metafeatures whose value is 0 or NA, have been included even though they were not initially to be considered. There are two reasons for this: firstly, the possibility exists that some time series classified as coming from households are actually not, and vice versa. This is particularly likely to be the case for the Spanish electric cooperatives dataset, as not all the CNAE codes provided (see Section 3.1.1) correctly identify households. The same applies for the Elergone Energia dataset (see Section 3.1.4), which does not provide explicit data on the origin of the time series. And secondly, the electrical behaviour patterns of commercial, industrial or any other non-domestic installations are expected to be easily recognisable.
- Second, time series with a percentage of imputed samples equal to or greater than 5% have been excluded. It is an indicator that the original time series had large gaps or too many scattered missing samples.
- And third, regarding the Spanish electric cooperatives dataset, time series having a tariff different from type “2.0TD” or equivalent (indicated by the `ref_atr_tariff` metafeature) have been excluded, as they belong to industrial or commercial sites.

After applying these criteria, 317 more time series have been excluded from the pool, resulting in the selection of 23 578 time series for the analysis.

6.1.2. Selection of cluster analysis method

Initially, the nine proposed cluster analysis methods were run on several datasets separately to test their clustering efficiency from a subjective point of view, although based on experience and supported by the validation measurements. Four groups can be established depending on the reliability of the clustering. From lowest to highest:

1. Clustering with no results. This is the case for the FANNY algorithm. It was not possible for the method to produce a valid result in an acceptable time. Tuning of certain parameters, such as the membership exponent, was tried, but without success.
2. Clustering generates a small group of well-populated clusters and a large group of sparsely populated clusters (outliers). This is the case for DIANA and SOTA, in which the well-populated clusters were around one third of the total; and Hierarchical, in which there was just one well-populated cluster.

3. Clustering is acceptable but some elements appear repeated or not well defined. This is the case of CLARA and the model-based clustering.
4. Clustering is good. The clusters appear without repetition and without noticeable issues. *k*-means, SOM, and PAM are in this group.

The three methods in the last group provide similar results but the SOM algorithm provides more identifiable clusters in the eyes of the domain experts. For this reason, the SOM algorithm has been chosen to be used in the rest of the project.

6.1.3. Selection of features

The large set of features extracted from the time series cannot be used directly for cluster analysis. The application of cluster algorithms to high-dimensional data often presents a series of issues, since they rely on detecting areas where objects form groups with similar characteristics, and high-dimensional data often appear to be sparse and heterogeneous. This is part of a well-known phenomenon called the *curse of dimensionality*[31].

The alternative is to select reduced subsets of similar features that can be organised in a meaningful way. Five sets of features have been designed for the application of cluster analysis methods. These sets are described below:

1. **Set of ADR seasonal aggregates**: contains the 24 features of the seasonal aggregates Group VIII, in its ADR version, related to the mean, and expressed as percentages. The `weekday` feature of the seasonal aggregates Group V, also in its ADR version, related to the mean, and expressed as a percentage, is also included in this set (25 features).
2. **Set of electric tariffication seasonal aggregates**: incorporates the same 25 features as in the previous set but taken from the seasonal aggregates Group XI.
3. **Set of peak and off-peak periods**: contains the peak and off-peak periods of Group I, the peak and off-peak periods of Group VI, and the peak period, divided by the number of days, of Group V (5 features).
4. **Set of seasonal strengths and autocorrelations**: contains the mean (from statistical moments); the entropy and the seasonal strength of days, weeks and years (from the `tsfeatures` package); and the correlation between values of the time series that are 1, 7 and 28 days apart (from lag *k*-day autocorrelations) (8 features).
5. **Catch-22 set**: includes the twenty-two *Catch22* features.

The reports of results of all these experiments can be found in the **Zenodo repository** entitled '*Analysis Reports on "Time Series from Smart Meters" Dataset*'[32]. All results were shown to domain experts that perform a judgemental evaluation of the results. Unanimously, the set of features that provided the most identifiable cluster visualisations was the set of **ADR seasonal aggregates**. For this reason, this set of features has been chosen to be used in the rest of the project.

[31] https://en.wikipedia.org/wiki/Curse_of_dimensionality
[32] https://zenodo.org/record/6344956

6.1.4. Selection of the number of clusters

Each cluster analysis method has been tested for 5, 10, 15, 20, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 50, 60 and 70 clusters. The resulting plots constructed from the validation measures have provided hints to narrow the assessment to the most promising values. Then, the domain experts have analysed the clusters found to decide the optimal number of clusters for each dataset.

The optimal number of clusters found for each working dataset are:

- 20 for the Spanish electric cooperatives;
- 16 for Low Carbon London and ISSDA;
- 6 for Elergone Energia; and
- 40 for the combination of all working datasets.

These 40 consumption patterns are analysed in Section 6.2.1 and can be found in the Zenodo repository[33].

The fact that each dataset has a different number of clusters may be due to a multitude of reasons: the size and type of the residences of the sampled households, their socio-economic conditions, their consumption habits, etc. This can make the number of consumption patterns vary from dataset to dataset. In addition, the number of time series of each dataset is also a relevant factor. Thus, it is not surprising that datasets with more samples have a larger number of patterns.

6.2. Analysis of results

Once the best combination of clustering method, set of features, and number of clusters per dataset has been determined, the strategy has focused on obtaining a set of results that answer three fundamental questions:

1. which are the main patterns of electrical consumption common to all the datasets analysed;
2. which are the differences in electricity consumption between regions;
3. and, which are the differences in electricity consumption, if any, between the pre-COVID-19 lockdowns stage (until the first quarter of 2020) and the post-lockdowns stage (from the second quarter of 2020).

6.2.1. The top 40 clusters

To obtain the main patterns of electrical consumption common to all datasets selected, the features of the "ADR seasonal aggregates" group have been clustered using SOM with forty centres (clusters). The means of all time series in the forty clusters have been plotted as heatmaps; they are available in **Annex D**.

[33] https://zenodo.org/record/6344956

Most of the heatmaps obtained provide a very detailed picture of the electricity consumption of households, which in turn corresponds reliably to their activities and behaviour. Other heatmaps depict what appear to be offices, while others present sunlight-dependent patterns. In order to classify and categorise the heatmaps in an efficient way, a taxonomy has been developed to organise the power consumption patterns by type. The final taxonomy has been achieved by combining two approaches: one based on expert knowledge/citizen science activity, and another based on an automatic hierarchisation.

The taxonomy based on expert knowledge/citizen science has been developed by professionals with knowledge on the electricity system. To assist in the process, the heatmaps have been printed in colour and cut out as cards, which have been manually grouped according to their characteristics (see Fig. 6).

Figure 6: Electric consumption patterns, in the form of heatmap cards, being manually grouped through expert knowledge.

The taxonomy based on an automatic hierarchisation has been generated through a **dendrogram**, obtained by hierarchical clustering. Using this technique, the visualisation matrices, and therefore the consumption patterns they represent, are categorised into a hierarchy similar to a tree-like diagram. The methodology applied is as follows: first, the values of the 40 visualisation matrices are rearranged into a vector of 8 904 elements and stored as a 40 x 8 904 matrix. The **distance matrix** between the elements (the vectorised visualisation matrices) is calculated using a Euclidean distance measure. This generates a dissimilarity structure: a triangular matrix providing the distances between each element of the matrix. The smaller the distance between two matrices, the more similar they are. A **hierarchical cluster analysis** using the agglomerative method can then be applied to the distance matrix, as performed by the R function `hclust()`, and the results are presented in a dendrogram.

The two taxonomies obtained turned out to be surprisingly similar. A further evaluation was carried out to unify both taxonomies, resulting in the one shown in **Annex E**. One of the advantages of the taxonomy is that it facilitates the navigation through its branches from the main node to any of the consumption patterns. By setting an appropriate question at each node, it is possible to discriminate between each of the child branches and arrive at the electric consumption pattern that characterises a particular consumer.

The questions have been merged, giving rise to the questionnaire in italics below. With a minimum of 3 questions and a maximum of 6, it is possible to identify, a priori, the electricity consumption pattern of any user[34]:

Question 1. *The electricity supply you have contracted with GoiEner, to which type of dwelling does it supply energy? If you have more than one contract, select one and fill in the rest of the section considering only that one.*

- *Main residence*
- *Second residence*
- *Small business, office or similar*
- *Common areas of buildings / External lighting*
- *Other (specify)*

[If the answer to Question 1 is "Main residence"] *Is someone at home all day on weekdays?*

- *Yes, one or several people teleworking*
- *Yes, one or several people doing housework*
- *Yes, one or several people are unemployed and/or retired (or similar status)*
- *No, we all work/study/spend the day out*

[If the answer to Question 1 is "Main residence"] *Is the house inhabited on weekends?*

- *Yes, and the same pattern is followed as during the week.*
- *Yes, but the pattern of behaviour changes*
- *No, we go to the second residence, visit relatives or others.*
- *Others*

[If the answer to Question 1 is "Main residence" or "Second residence"] *In what range of hours are the following actions performed in the household? Please include the range from when the first household member performs them until the last household member finishes. In shift households, try to follow cultural conventions. Finally, if an action is not performed in the household, please leave it blank.*

	Start time on weekdays	*End time on weekdays*	*Start time on weekends*	*End time on weekdays*
Wake up				
Breakfast				

[34] A validation activity is planned to be carried out latter in the project and will be reported at Deliverable D2.3

Lunch				
Dinning				
Sleep				

[If the answer to Question 1 is "Second residence"] *What time of the year do you live there? (More than one answer is possible)*

- *Every weekend or most weekends*
- *Long weekends*
- *Autumn*
- *Winter*
- *Spring*
- *Summer*

[If the answer to Question 1 is "Small business, office or similar"] *What are your opening hours?*

- *Morning only (8:00 to 15:00 or similar)*
- *Morning and afternoon split timetable (9:00 to 13:00 and 17:00 to 20:00, or similar)*
- *Morning and evening continuous hours (9:00 to 21:00 or similar)*
- *Other*

Question 2. *What is the main type of heating used in your home?*

- *Individual gas boiler*
- *Electric boiler*
- *Energy accumulators*
- *Electric radiators*
- *Heat pump (Aerothermal, Air Conditioning or other)*
- *Butane*
- *Oil boiler*
- *Biomass heater or boiler*
- *Building central heating*
- *Heat network / District heating*
- *I live in a passive house standard building*
- *We don't have a boiler / We don't use it*
- *Don't know*

Question 3. *Do you have any of the following electrical systems in your household? (Check all that apply)*

- *Batteries*
- *Electric boiler / Heat pump (for domestic hot water)*
- *Electric cooker*
- *Electric vehicle*
- *Other*

Alongside the taxonomy, a labelling of the forty patterns has been established through expert knowledge. In addition, as the groups are very well defined, a *persona* of each household has been developed, i.e. a narrative description of the possible household type behind each pattern. Labels and personae can be found in **Annex D**.

6.2.2. Regional comparison assessment

Based on the clustering carried out in the previous section, the country composition of each cluster has been analysed, making it possible to perform a regional comparison.

For this purpose, for each country only those consumption patterns with more than 5% of the load profiles have been represented as a bar plot (see Fig. 7). Each colour bar indicates both a different pattern and its proportion with respect to the rest of the patterns within each country.

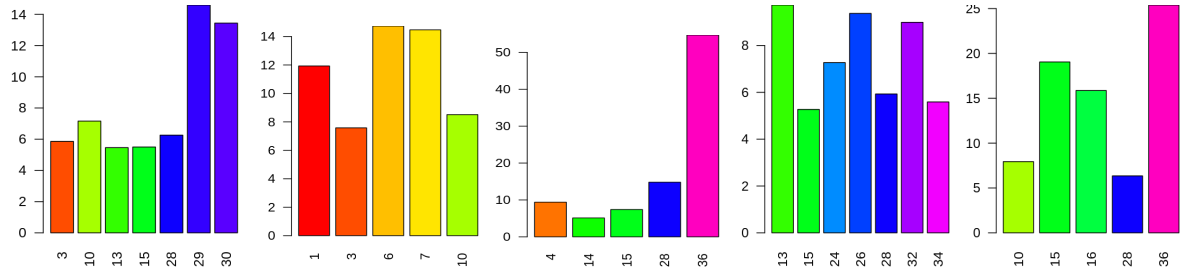


Figure 7: Clusters with more than 5% of the time series, from left to right: ES, GB, IE, PT, US.

The assessment provides the following highlights:

- First, two clusters are among the most prevalent in all regions except Ireland. These are the patterns numbered #15 and #28 (see Fig. 8). These consumption patterns are associated with regular, all-day-at-home behaviour. The main differences between #15 and #28 are that a) #15 has less consumption in summer and #28 shows no change in summer and b) #28 has a strongest week-weekend patterns showing a larger energy consumption on weekends than on weekdays where in #15 this trend is not so obvious.

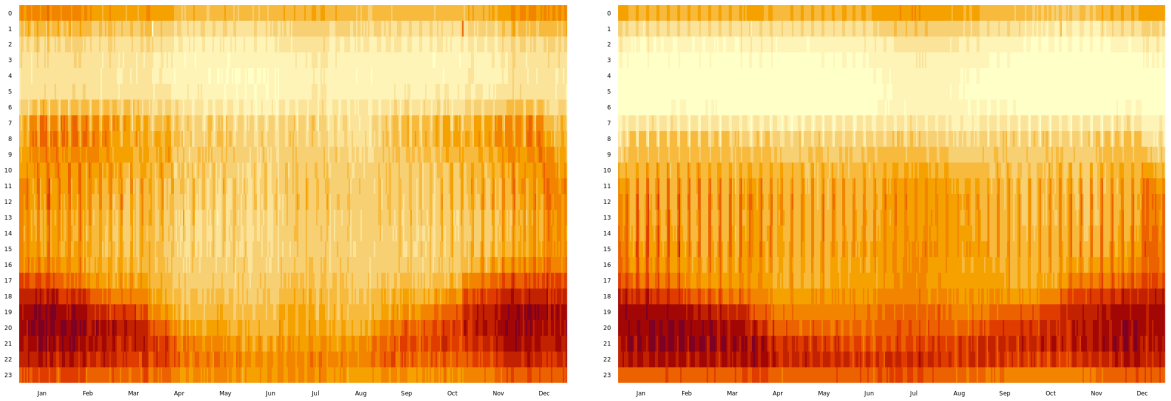


Figure 8: Consumption patterns #15 (left) and #28 (right). See Annex D for further details.

- Second, the Irish dataset is the one that shows the greatest differences with respect to the rest, although it shares many characteristics with other Anglo-Saxon countries analysed, mainly the UK and the US.
- Lastly, Ireland and the US share a pattern that clearly dominates the dataset. This is pattern #36 (see **Annex D**), which contains small and medium-sized companies that do not close on holidays and open from 6 am. In particular, the five most popular clusters comprise more than 90% of the time series in Portugal and the US (in the rest, it is 60%).

Please note that this assessment has some limitations: first, there is a lack of random sampling. Dataset collected are based on volunteers or customers, so the results must be taken cautiously and they should not be used to generalise. Second, data lacks context in many cases, so it is not easy to stratify further and correct for potential biases in order to extrapolate the conclusions. Finally, it is important to remark that we are analysing some datasets with small amounts of load profiles of which their origin is unknown, as is the case of the Portuguese (*Elergone Energia*) dataset.

6.2.3. Pre- & post-COVID-19 lockdowns assessment

To analyse the impact that the COVID-19 pandemic has had on energy consumption habits, the load profiles provided by the Spanish electric cooperatives have been split into two sub-datasets (see Section 3.2.3) and clustered independently. Note that the objective of this assessment is not the lockdown itself but the change in behaviour that occurs.

The first sub-dataset (**pre-COVID-19 dataset**) contains all time series with over one year of electricity records that avoid the time period comprising the Spanish COVID-19 lockdowns, from March 14 to May 9, 2020. To achieve this, all time series in this sub-dataset have March 1, 2020 as the cut-off date. The second sub-dataset (**post-COVID-19 dataset**), comprises all time series with over one year of information recorded from March 1, 2020 onwards. Because the Spanish electric cooperatives have been gaining customers over time, the 'post' dataset contains more time series (16 359) than the 'pre' dataset (12 189).

In this case, since we have only used a subset of the data used for the creation of the top 40 clusters (see Section 6.2.1), we decided to reduce the number of clusters to 20. Then, a mapping process between 'pre' and 'post' clusters has been carried out using as the connection point the already existing top 40 clusters. Therefore, each of the 20 clusters of both the 'pre' and 'post' intervals have been linked to one of the top 40 clusters. To do so, a semi-automatic approach has been used using the Euclidean distance between clusters and choosing the nearest one as the match. Nevertheless, all the automatically matched pairs have been manually validated since, in some cases, cluster distances were very similar.

The final result of this matching process is shown in Fig. 8. The first worth mentioning aspect of this process is that 17 out of 20 generated clusters with pre-COVID-19 load profiles have been found in the previously introduced 40 clusters. The missing three clusters have been considered outliers since they only represent 2.19% of the total

dataset. When it comes to the post-COVID-19 clusters, 19 out of the 20 generated clusters have a match, the missing cluster representing 0.51% of the total dataset.

Comparing which clusters are missing in the 'pre' dataset with respect to the 'post' dataset and vice versa, first we see that cluster #9 has not been found in the 'post' dataset. This cluster represents the use of electric heating in winter, which is turned down during working days. Therefore, it may make sense that with the increase in teleworking the electric heating is not turned down on weekdays and consequently this cluster may have become less significant.

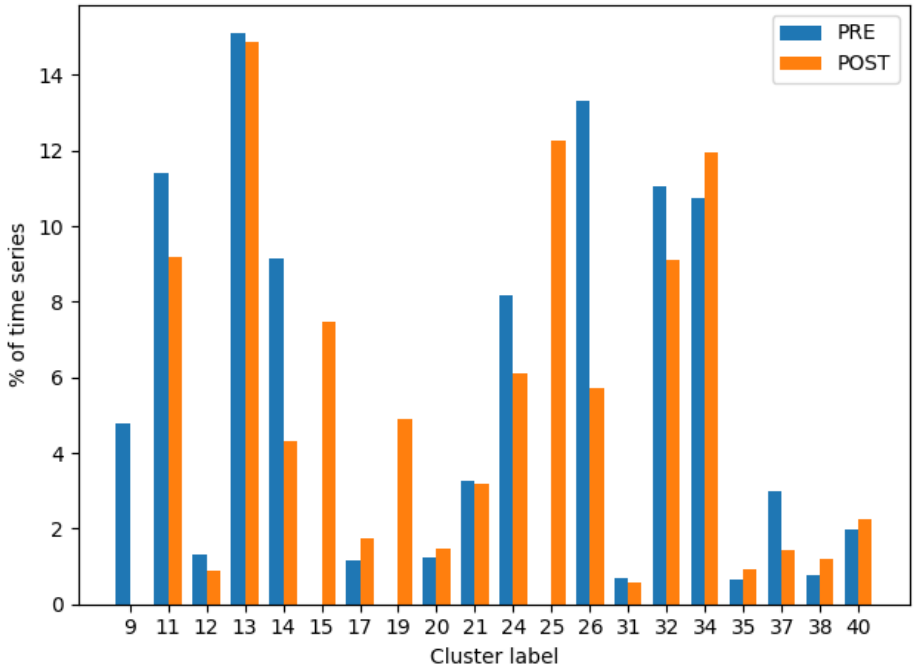


Figure 8: Distribution in terms of percentage of the time series/load profiles of the PRE and POST datasets among the selected clusters.

On the one hand, with regard to the clusters that are missing in the 'pre' dataset but we have identified in the 'post' dataset, cluster #19 (electric heating used in winter all day) is missing on the 'pre' dataset, whereas a match has been found in the 'post' dataset. In this case, the reasoning behind this fact might be very similar to what has happened with cluster #9: the use of electric heating during all day may have increased due to the pandemic lockdowns.

On the other hand, 12 out of the 20 missing clusters belong to specific patterns of consumption in households (inhabitants only have dinner or breakfast, they are at home all the day but most of the consumption is in the evening, etc.). Therefore, most of these specific patterns inside a household might have been unified in some of the 20 clusters we have used in the 'pre' and 'post' experiments or are not represented in the dataset provided by the Spanish electric cooperatives. Apart from that, we can see that two clusters where the electric heating is the main source of the energy consumption have disappeared (#2 and #8), clusters representing the load profile of businesses (#4, #22, #36) that are open all day and all year round are also missing, or cluster #39 (external lighting from a neighbourhood community) have not been detected either.

Focusing on the differences in terms of how the 'pre' and 'post' load profiles have been distributed among the clusters, the first worth mentioning aspect is the difference on cluster #14 (second homes for holidays and weekends) from 'pre' (9.14%) to 'post' (4.31%), meaning that visits to the second homes have been reduced drastically (strict mobility restrictions were applied in Spain due to the COVID-19). This phenomenon also occurs with cluster #37 (second homes visited almost every weekend of the year and on holidays) with a decrease of 1.5 percentage points.

A similar pattern can be seen with cluster #26, in which a reduction of more than 7 percentage points can be identified. This cluster represents a household where people work away from home and most of their energy consumption occurs in the evenings and change their behaviour in the weekends and holidays. Again, this could also be due to the changes made to the Spanish lifestyle due to the pandemic. Something similar occurs with #11 (people working away) showing a decrease of 2 percentage points with respect to the 'pre' dataset.

Conversely, the importance of clusters such #34 (energy consumption during all day but with a behaviour change on weekends) has been increased by more than 1 point in the 'post' dataset.

Apart from the above, after visually analysing the heatmaps generated by each of the 'post' clusters, some indicators of the most strict lockdowns in Spain can be easily identified. For instance, in Fig. 9 (left) it can be seen when the first lockdown started in Spain (mid-March) and how mobility restrictions affected the first days of December, which is a period with contiguous bank holidays. In Fig. 9 (right) this phenomenon is even more visible.

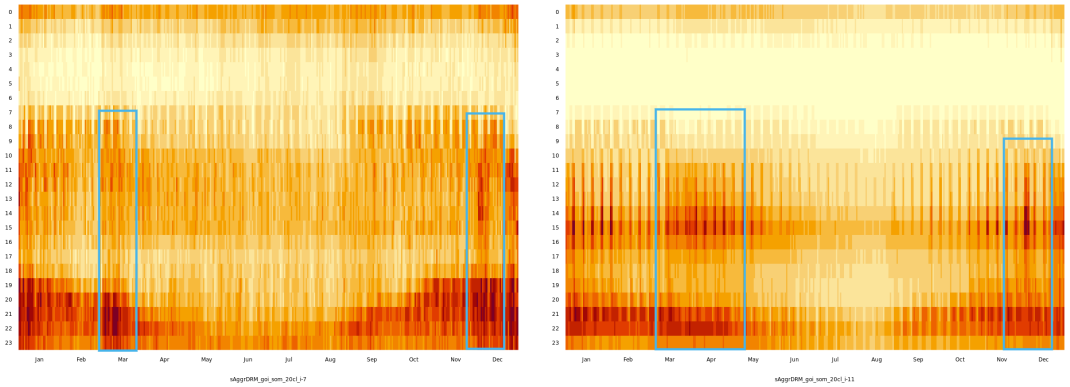


Figure 9: Heatmaps showing the energy consumption on lockdowns for clusters #15 (left) and #34 (right).

7. Future work and exploitability potential

The assessments carried so far are extensive but have raised a lot of new questions. In this section we plan to present some of the assessments we would like to carry out in the short term and other assessments that could be exploited later on (by the partners of the project or other interested stakeholders).

7.1. Methodology

Here we present all the ideas related to the features to use or the methodology to construct the universe of clusters. In particular, these ideas has been discussed but not put into practice so far:

- The exact location of each time series could not be published in order to avoid re-identification of the time series. In order to mitigate the loss of socio-economic and climate driven contextual information, we plan to include new socio-economic and climate metafeatures based on the exact location in categories that reduce or avoid the re-identification.
- Include as features the correlation of the time series with the weather variables such as temperature, solar radiation, etc. and incorporate these values as new features.
- Make a taxonomy of all possible day types in terms of energy consumption following a similar approach as used so far (but or days, instead of years) and calculate the percentages of each day type for each time series. Incorporate the results as new features.
- Apply model reduction techniques (such as PCA) to extract the principal components, i.e. a linear basis, of each consumption pattern. Use this technique to deduce automatically the pattern to which a time series belongs.
- Validate that the instrument to classify a time series in one of the clusters works as expected. This is expected to be done later on in the project and the results will be provided in Deliverable D2.3.

7.2. Regional differences

The regional assessment carried out so far is quite simple given the data we currently have. More sophisticated assessment could be made if Here we present all the ideas related to the features to use or the methodology to construct the universe of clusters. In particular, these ideas has been discussed but not put into practice so far:

- Repeat the methodology for each dataset individually and compare the resulting cluster to check if new clusters appear. In some cases, the amount of data does not allow us to make this kind of assessment as the data is highly unbalanced.
- Check a methodology to select the best number of clusters at regional and global level by implementing an algorithm that modifies both the numbers of clusters at regional and global level and then compute the association error between each regional cluster centre with the nearest global cluster.

7.3. Lockdown impacts

- Comparison of the number of time series that have changed their cluster type pre- and post lock-downs and what were the changes. A different methodology than the one used so far will be needed. The most sensible one is to classify the post lock-down features to the nearest neighbourhood and compute a "confusion" matrix[35].
- Assess the social impact of the lock-downs in terms of what cluster types have the largest changes in energy consumptions. For this, it is needed to define a set of KPIs to assess (total energy consumption on the lock down, for example) and compute for each one of the clusters. Assess the biggest difference between cluster types taking into consideration the socio-economic description of the clusters.
- Assess if a change in a tariff or a communication campaign could have an impact as severe as a lockdown. Basically, by comparing the results of this deliverable with the relevant results of D2.3.

7.4. Exploitation by partners and stakeholders

- Evaluate results jointly with other sister projects of the same call, such as NewTrends. This will allow an increase in the amount of countries considered and the type of assessment carried out.
- Triangulate the results with initiatives from other stakeholders like the SPAHOUSEC III from IDAE[36] or HETUS from EUROSTAT[37]. If successful, this will allow to generalize the results found as these initiatives follow better sampling methods than the one used in this project.

[35] Note that this will not be a confusion matrix as there is not a correct model.
[36] https://www.idae.es/en/node/23156
[37] https://ec.europa.eu/eurostat/web/time-use-surveys

ANNEX A: Technical characteristics of data extraction

This Annex contains technical information on how data extraction (see Section 3.1) is performed per dataset. Detailed procedures can be found for the *Spanish electric cooperatives*, *ISSDA*, and *Low Carbon London* datasets. Metadata extraction is also discussed.

A.1. Spanish electric cooperatives

A.1.1. Specificities of the Spanish electric market

GoiEner provides the data of electricity consumption (and generation) for every CUPS in its database. A **CUPS** (*Código Unificado de Punto de Suministro*, or "Unified Supply Point Code" in English) is a code of 20 to 22 alphanumeric characters used in Spain to uniquely identify electricity supply points, such as individual domestic residences or businesses. It should be noted that, prior to any file processing, all CUPS in the dataset have been automatically replaced by unique 32-digit hexadecimal MD5 hash codes[38] in order to make them unrecognisable and thus comply with the EU laws (e.g. GDPR[39]) on the protection of personal data.

The files in the GoiEner dataset follow the format described by SIMEL (*Sistema de Información de Medidas Eléctricas,* "Electrical Metering Information System" in English). SIMEL is a platform for the exchange of information among all the stakeholders involved in the Spanish energy system. It describes the different formats of the files for the exchange of information used by the participants of the measurement system. The format in which the measurements are stored in the files is determined by their file type.

As of November 2021 (version 37), SIMEL documentation[40] describes about 140 different file types for managing the exchange of information among the different actors. However, the bulk of files in the dataset provided by the Spanish energy cooperatives is stored in only a dozen different file types, as shown in Fig. A1.

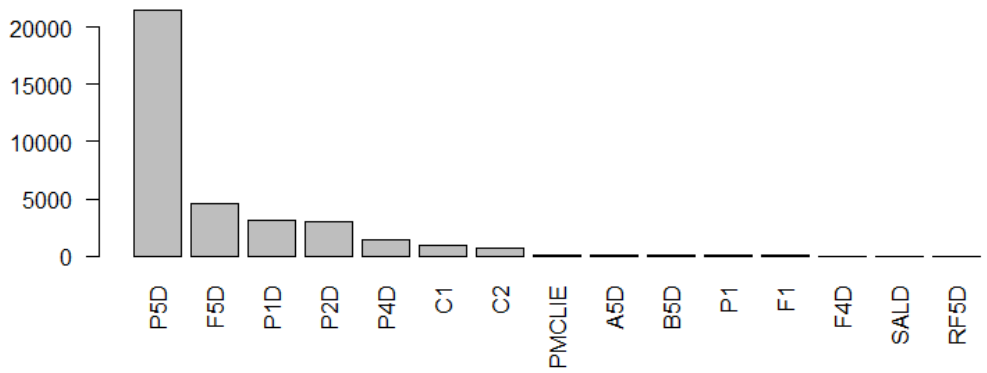


Figure A1: Number of SIMEL files by file type

[38] https://en.wikipedia.org/wiki/MD5
[39] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[40] Sistema de Información de Medidas Eléctricas, *Ficheros para el intercambio de información de medida* (Versión 37, noviembre de 2021), Dirección General de Operación.

The file type is identified by the first characters of the file name. For the SIMEL files in the dataset, the file name usually has the following format:

`type_codes_date.v`

where:

- **`type`**: is the file type (P5D, F5D, P1D, P2D, ...). Each file type is associated with a way of communicating electricity data, i.e. how fields are formatted. The file type also depends on the type of customer (see below for more details).
- **`Identification codes`**: is, in general, a series of 4-digit codes referring to the electricity distributors and/or retailers, although it varies depending on the file type.
- **`dDate`**: is the date of creation of the file in the format `yyyymmdd`.
- **`v`**: is the file version. If the information is published in several files per day, consecutive versions (with the same name) are used, starting with version 0.[41]

Some examples of SIMEL filenames are shown below:

- `P5D_0021_1377_20200201.2`
- `F5D_0031_0856_20200818.0`
- `C1_0021_1139_20200610.0`

A.1.2. SIMEL files

SIMEL files are stored as plain text (ASCII) and are structured as a sequence of text lines. Each line contains a number of fields separated by semicolons (;), which varies depending on the file type. Line breaks are also preceded by semicolons. Fields are identified by a letter, starting with A and following the alphabetical order. Fields that do not contain data are left empty.

Not all file types and not all fields are relevant for the extraction of information on electricity consumption (or generation). Table A1 summarises the relevant file types and their fields:

File type	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
A5D	C	T	F	IN [Wh]						DCM											
B5D	C	T	F	IN [Wh]	OUT [Wh]	Q1 [VArh]	Q2 [VArh]	Q3 [VArh]	Q4 [VArh]	DCM											
F1	C		T	F	IN	OUT	Q1	Q2	Q3	Q4			DCM								
F5D	C	T	F	IN [Wh]	OUT [Wh]	Q1 [Wh]	Q2 [Wh]	Q3 [Wh]	Q4 [Wh]	DCM											

[41] As described in the official document '*Resolución de 2 de junio de 2015, de la Secretaría de Estado de Energía, por la que se aprueban determinados procedimientos de operación para el tratamiento de los datos procedentes de los equipos de medida tipo 5 a efectos de facturación y de liquidación de la energía*'.

P1D	C		T	F	IN		OUT		Q1		Q2		Q3		Q4					DCM
P2D	C		T	F	IN		OUT		Q2		Q3		Q4		Q5					DCM
P5D	C	T	F	IN [Wh]	OUT [Wh]															
RF5D	C	T	F	IN [Wh]	OUT [Wh]	Q1 [Wh]	Q2 [Wh]	Q3 [Wh]	Q4 [Wh]	DCM										

Table A1: Summary of the most relevant fields and SIMEL file types.

Some remarks about the table:

- Fields highlighted in red with a letter 'C' contain the original *Universal Supply Point Code* (CUPS).
- Fields highlighted in orange with letters 'T' and 'F' contain, respectively, the date and time of the measurement in the format '`yyyy/mm/dd hh:mm`', and a binary summer/winter flag indicating whether the daylight saving time is on for that time.
- Fields highlighted in yellow with the label 'IN' contain the 'measurement of the incoming active value', i.e. the energy consumed by the CUPS in an hour. The default units are kWh, unless otherwise stated.
- Fields highlighted in green with the label 'OUT' contain the 'measurement of the outgoing active value', i.e. the energy generated by the CUPS in an hour. The default units are kWh, unless otherwise stated.
- Fields highlighted in blue with the label 'DCM' contain the 'data collection method'. It can take any value between 1 and 6, with 1-3 being firm measurements and 4-6 being provisional measurements. As a rule of thumb, the lower the value of this field the more reliable the measurements.
- Fields with labels 'Q1' to 'Q4' are also indicated, containing the 'measurement of the reactive values' in each of the four quadrants. The default units are kvar, unless otherwise stated. These values have not been used at this point in the Project.

As for the file types, some of them are described in the SIMEL documentation as:

- **F1**: communication of hourly data of energy at a boundary point of type-1 and type-2 customer.
- **F5D**: publication of hourly data of incremental energy at a boundary point of type-5 customer.
- **P1D**: communication of hourly data of energy at a supply point of type-1, type-2 and type-3 customer and self-consumption of type-4 customer.
- **P2D**: communication of quarter-hourly data of energy at a supply point of type-1 and type-2 customer.
- **P5D**: communication of hourly data of validated gross incremental energy at a boundary point of type-5 customer.
- **RF5D**: communication of hourly data of incremental energy at a boundary point of type-5 customer after a complaint.

Regarding the file types **A5D** and **B5D**, they are provisional files used to report self-consumed hourly energy. Although they do not yet appear in the SIMEL documentation because they are pending approval, their description can be found in the

documents provided by the Spanish National Commission for Markets and Competence (CNMV).

Finally, the file types that have not been mentioned but appear in Fig. A1 (P4D, C1, C2, PMCLIE, P1, F4D, and SALD) either have a very low occurrence or the data they provide are not relevant. They have therefore been discarded during processing.

A.1.3. File processing scheme

The process by which the raw files (CSV files with timestamped consumption values) are obtained from the SIMEL files is achieved in three steps:

1. Processing of SIMEL files to obtain CUPS files.
2. Correction of repeated timestamps in CUPS files.
3. Processing of CUPS files to obtain the raw files.

A single SIMEL file may contain energy usage data for one or more CUPS. Therefore, it is convenient to generate a single file per CUPS (i.e. the **CUPS files**) that gathers, from all SIMEL files, all the data related to each CUPS (as illustrated in Fig. A2). To achieve this, all lines with a particular CUPS in any SIMEL file are copied to its corresponding CUPS file. In order to know which file each line is extracted from (and thus have some traceability), the name of the original SIMEL file is prepended as a new field to the copied line. As a result, a collection of files is obtained, each with all its lines belonging the same CUPS. As CUPS files are made up of lines from different SIMEL files, these lines can have different formats.

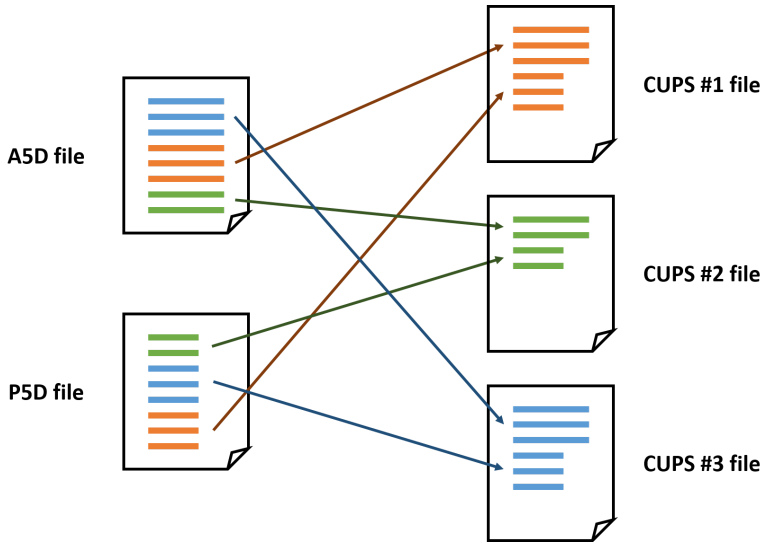


Figure A2: SIMEL files (left) are processed to generate files with unique CUPSs (right). Lines with different CUPS are indicated with different colours and lines with different formats are indicated with different line lengths.

The next step consists in identifying, for each CUPS file, all lines (if any) with identical timestamps, i.e. with the same dates and times. There are at least two reasons why lines with repeated timestamps may appear in CUPS files. The first is because of the SIMEL files with complementary information, such as A5D and B5D file types, which provide data on

consumption and generation. The second possible reason appears when more reliable measurements are obtained and the data is updated (e.g. there has been a loss of communication between the smart metre and DSO. In these cases it is usually sent an estimation to the energy retailer but when the issue is solved, a correction could also be sent).

Since it has been determined that raw files can only have non-repeating timestamps, it is necessary to disambiguate all repeated timestamps that may appear in the CUPS files, especially when they report different values of incoming (or outgoing) energy. The criteria established to correct these mismatches are as follows:

1. If the incoming (or outgoing) energy values are the same in all repeated entries, then that value is considered as correct.
2. If the incoming (or outgoing) energy values are different in at least one repeated entry and the field containing the 'data collection method' field exists in all entries, then the value that prevails is the one corresponding to the lowest 'data collection method' value, since it is the most reliable one.
3. If the incoming (or outgoing) energy values are different in at least one repeated entry and their file types are F5D, A5D or B5D, then the incoming energy value that prevails is calculated as the sum of all incoming energy values; and the outgoing energy value that prevails is calculated as the sum of all outgoing energy values. This is because A5D and B5D encode consumed self-generated energy, which has to be added to the consumed energy from the grid encoded in F5D.
4. If the incoming (or outgoing) energy values are different in at least one repeated entry and their file types are P5D or F5D, then the value that prevails is that of P5D, since it is a validated value (as opposed to that of F5D, which is a billing value[42]). Note that no 'data collection method' field exists for P5D.
5. Otherwise, discard the CUPS file. This case does not arise in practice because all cases are covered by the above conditions.

Finally, once the timestamps of all CUPS files are unique and sorted in ascending order (most recent last), the generation of raw files is straightforward, by taking the values of the corresponding fields according to the format of each line. When values of generated electricity are obtained in addition to those of consumed electricity, they are placed as a third column in the raw file. Missing data are imputed in a later data cleaning stage (see Section 3.2.1).

A.1.4. Metadata files

Metadata is provided as a unique CSV file which stores, for each record:

- The CUPS, duly replaced by a MD5 32-character hash code to protect the identity of customers. Customers include not only households but also businesses, companies, offices, and public facilities.
- Start and end (if any) date of contract of the customers with the electricity company, as well as the characteristics of the tariffs and contracted powers.

[42] As described in the official document '*Resolución de 2 de junio de 2015, de la Secretaría de Estado de Energía, por la que se aprueban determinados procedimientos de operación para el tratamiento de los datos procedentes de los equipos de medida tipo 5 a efectos de facturación y de liquidación de la energía*'.

- Municipality, province, and post code in which the CUPS is registered.
- CNAE 4-digit code of the registered CUPS. CNAE stands for 'National Classification of Economic Activities' (from Spanish '*Clasificación Nacional de Actividades Económicas*'), and classifies and groups the production units according to the activity they carry out for the purpose of compiling statistics. This makes it possible to identify (in theory) those CUPS belonging to households.

There may be several records for the same CUPS, indicating changes in the contract. During data processing, the information from the most recent contract has been considered, while keeping the oldest customer registration date.

A.2. Irish Social Science Data Archive (ISSDA)

A.2.1. File format

Smart metre read data is stored in six zipped files named 'File1.txt.zip' to 'File6.txt.zip'. Each file contains a large text file with several tens of millions of space-separated records. There are three fields per record, corresponding to:

- a 4-digit value identifying the smart metre;
- a 5-digit value encoding the timestamp (the first three digits encode the day, where 1 is 1 January 2009; whereas the last two digits encode the time interval, which is between 1 and 48, where 1 is the period between 00:00:00 and 00:29:59).
- the value of electricity consumed during the 30 minute interval (in kWh).

The processing of the files is straightforward, simply by separating the entries into different files according to the smart metre identifier. Timestamp decoding and subsequent sorting of the time series are also required.

A.2.2. Metadata files

Some files exist that provide metadata: the allocation file and a series of pre- and post-trial surveys for residences and SMEs.

The allocation file identifies the load profiles by their 4-digit ID value, specifying if they belong to households, SMEs or other categories. It also provides information on the residential stimulus and tariffs.

Pre and post trial surveys contain coded computer-assisted telephone interviews (CATI) conducted with household or SME owners. The questionnaires can be freely accessed from the ISSDA website[43]. The questions focus both on personal and behavioural data of the respondent and their household as well as on the characteristics of their dwelling from the

[43] https://www.ucd.ie/issda/static/documentation/cer/smartmeter/cer-residential-pre-trial-survey.pdf
https://www.ucd.ie/issda/static/documentation/cer/smartmeter/cer-sme-pre-trial-survey.pdf
https://www.ucd.ie/issda/static/documentation/cer/smartmeter/cer-residential-post-trial-survey.pdf
https://www.ucd.ie/issda/static/documentation/cer/smartmeter/cer-sme-post-trial-survey.pdf

point of view of energy consumption. Only a subset of questions of the residential pre-trial survey has been incorporated into our analyses (see **Annex C**).

A.3. Low Carbon London

A.3.1. File format

The dataset contains 112 CSV files with three fields of information: an alphanumeric ID that uniquely identifies each household, the timestamp of date and time, and the energy consumption in kWh (per half hour). The CSV file is around 10GB when unzipped and contains around 167 million rows.

The processing of the files is straightforward, simply by separating the entries into different files according to the household identifier. Timestamp decoding and subsequent sorting of the time series are also required.

A.3.2. Metadata files

This dataset contains several files with metadata, the most relevant for our analysis being 'informations_households.csv'. This file specifies for each household the demographic group to which it belongs according to the Acorn segmentation.[44,45] Three large Acorn categories (*Affluent*, *Comfortable*, and *Adversity*) are established in this dataset, which are further subdivided into 18 more specific groups ranging from A to Q (see Fig. A3). In addition to this segmentation, the file also provides information on whether the households belonged to the group subjected to the '*Dynamic Time of Use*' (dToU) energy prices throughout the 2013 calendar year period or to the *Standard* energy prices.

The metadata is complemented by files providing hourly and daily weather during the studied period, a list with the UK bank holidays, and the distribution of other demographic and socioeconomic factors across the 18 Acorn groups, such as ethnicity, religion, level of qualifications, number of children, etc.

[44] https://acorn.caci.co.uk/
[45] https://acorn.caci.co.uk/

Category	Group	
1. Affluent Achievers	A	Lavish Lifestyles
	B	Executive Wealth
	C	Mature Money
2. Rising Prosperity	D	City Sophisticates
	E	Career Climbers
3. Comfortable Communities	F	Countryside Communities
	G	Successful Suburbs
	H	Steady Neighbourhoods
	I	Comfortable Seniors
	J	Starting Out
4. Financially Stretched	K	Student Life
	L	Modest Means
	M	Striving Families
	N	Poorer Pensioners
5. Urban Adversity	O	Young Hardship
	P	Struggling Estates
	Q	Difficult Circumstances

Figure A3: Table summarising the ACORN classification, partially extracted from the 'ACORN user guide'.

ANNEX B: Nomenclature of the seasonal aggregate features

The names given to all **seasonal aggregate features** follow this pattern:

`[abs|rel]_[mean|sd|sum]_groupName[|_pday|_drm]`

The first block of the name, **`[abs|rel]`**, indicates if the feature provides an absolute or a relative value. The second block, **`[mean|sd|sum]`**, indicates if the feature provides the mean, the standard deviation, or the sum, respectively, of the time bands in a Group. The third block, **`groupName`**, varies depending on the computed Group, shown below:

- **Group I**: **`XXh`**, where XX is a 2-digit value that can be 00, 01, 02, …, 23, expressing an hour in a 24-hour clock (e.g. **`18h`** refers to the "6 pm" time band).
- **Group II**: **`XXhYYh`**, where XX and YY are 2-digit values that can be 00, 04, 08, 12, 16, and 20, expressing hours in a 24-hour clock (e.g. **`16h20h`** refers to the "4 pm to 7 pm" time band).
- **Group III**: **`XXhYYh`**, where XX and YY are 2-digit values that can be 00, 06, 12, and 18, expressing hours in a 24-hour clock (e.g. **`00h06h`** refers to the "12 am to 5 am" time band).
- **Group IV**: the 3-letter lowercase abbreviations for the days of the week (e.g. **`thu`** refers to the "Thursday" time band).
- **Group V**: the keywords **`weekday`** (for the "weekday" time band) and **`weekend`** (for the "weekend" time band).
- **Group VI**: the 3-letter lowercase abbreviations for the months (e.g. **`dec`** refers to the "December" time band).
- **Group VII**: the keywords **`spring`** (for the "spring" time band), **`summer`** (for the "summer" time band), **`autumn`** (for the "autumn" time band), and **`winter`** (for the "winter" time band).
- **Group VIII**: **`XXhYYhZZZ`**, where XX and YY are 2-digit values that can be 00, 04, 08, 12, 16, and 20, expressing hours in a 24-hour clock, and ZZZ are 3-letter lowercase abbreviations for the seasons (e.g. **`04h08hspr`** refers to the "4 am to 7 am in spring" time band).
- **Group IX**: **`td2.0_p6_XXhYYh`**, where XX and YY are 2-digit values that can be 00, 08, 10, 14, 18, and 22, expressing hours in a 24-hour clock (e.g. **`td2.0_p6_18h22h`** refers to the "6 pm to 9 pm" time band).
- **Group X**: **`td2.0_p7_ZZZ`**, where ZZZ can be either (1) **`XXhYYh`**, where XX and YY are 2-digit values that can be 00, 08, 10, 14, 18, and 22, expressing hours in a 24-hour clock; or (2) the keyword **`wkends`** for the "weekend" time band (e.g. **`td2.0_p7_wkends`** refers to the "weekends" time band).
- **Group XI**: **`td2.0_p6_XXhYYh_ZZZ`**, where XX and YY are 2-digit values that can be 00, 08, 10, 14, 18, and 22, expressing hours in a 24-hour clock, and ZZZ are 3-letter lowercase abbreviations for the seasons (e.g. **`td2.0_p6_14h18h_aut`** refers to the "2 pm to 5 pm in autumn" time band).
- **Group XII**: **`td2.0_p7_ZZZ_WWW`**, where ZZZ can be either (1) **`XXhYYh`**, where XX and YY are 2-digit values that can be 00, 08, 10, 14, 18, and 22, expressing hours in a 24-hour clock; or (2) the keyword **`wkends`** for the "weekend" time band; and WWW are 3-letter lowercase abbreviations for the seasons (e.g. **`td2.0_p7_10h14h_win`** refers to the "10 am to 2 pm in winter" time band).

- **Group XIII**: **`td2.0_p3_XXX`**, where XXX can be the keywords **`punta`** (for the "punta" time band), **`valle`** (for the "valle" time band), and **`llano`** (for the "llano" time band).
- **Group XIV**: **`td2.0_pmy_XXX_YY_ZZ`**, where XXX can be the keywords **`punta`**, **`valle`**, and **`llano`**, YY can be a 2-digit value from 01 to 12 expressing the month, and ZZ can be a 2-digit value from 14 to 22 expressing the two last digits of the year (e.g. **`td2.0_pmy_llano_06_18`**, refers to the "*llano* in June 2018" time band).

The fourth block, **`[|_pday|_drm]`**, indicates if the feature has been computed as a TBR value (no suffix), as a TBR value per day (**`_pday`** suffix), or as an ADR value (**`_drm`** suffix). Remember that all features in Groups IX to XIV are computed as ADR values and, therefore, this is not indicated by any suffix.

ANNEX C: ISSDA survey answers metafeatures

The answers to the questions listed below are incorporated as metafeatures to the database of features.

From survey answers #1:

- 200: PLEASE RECORD SEX FROM VOICE.
- 47001: Do you have a timer to control when your heating comes on and goes off?
- 47011: Do you have a timer to control when your hot water/immersion heater comes on and goes off?
- 4801: Do you use your immersion when your heating is not switched on?
- 471: Returning to heating your home, in your opinion, is your home kept adequately warm?
- 473: Have you had to go without heating during the last 12 months because of a lack of money?
- 420: How many people over 15 years of age live in your home?
- 430: And how many of these are typically in the house during the day (for example for 5-6 hours during the day)?
- 4551: What rating did your house achieve? (1) A; (2) B; …; (7) G
- 43521: If you were to make changes to the way you and people you live with use electricity, how much do you believe you could reduce your usage by? (1) 0%; (2) <5%; (3) 5%-10%; (4) 10%-20%; (5) 20%-30%; (6) >30%.
- 4531: Approximately how old is your home? (1) <5 years; (2) <10 years; (3) <30 years; (4) <75 years; (5) >75 years.
- 453: What year was your house built?
- 6103: What is the approximate floor area of your home? (square metres)

From survey answers #2:

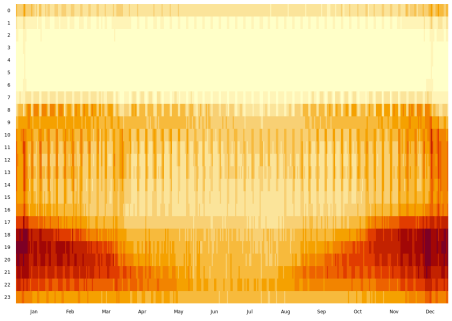
- 410: What best describes the people you live with? (1) I live alone; (2) All people in my home are over 15 years of age; (3) Both adults and children under 15 years of age live in my home.
- 4321: Multiple: (1) I/we have already done a lot to reduce the amount of electricity I/we use; (2) I/we have already made changes to the way I/we live my life in order to reduce the amount of electricity we use; (3) I/we would like to do more to reduce electricity usage; (4) I/we know what I/we need to do in order to reduce electricity usage.
- 450: I would now like to ask some questions about your home. Which best describes your home? (1) Apartment; (2) Semi-detached house; (3) Detached house; (4) Terraced house; (5) Bungalow; (6) Refused.
- 452: Do you own or rent your home? (1) Rent (from a private landlord); (2) Rent (from a local authority); (3) Own Outright (not mortgaged); (4) Own with mortgage, etc; (5) Other.
- 470: Which of the following best describes how you heat your home? (1) Electricity (electric central heating storage heating); (2) Electricity (plug in heaters); (3) Gas; (4) Oil; (5) Solid fuel; (6) Renewable (e.g. solar); (7) Other.

From survey answers #3:

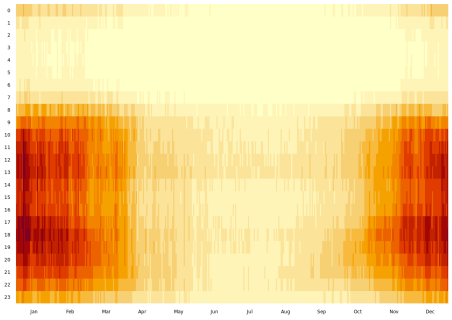
- 4701: Which of the following best describes how you heat water in your home? (1) Central heating system; (2) Electric (immersion); (3) Electric (instantaneous heater); (4) Gas; (5) Oil; (6) Solid fuel boiler; (7) Renewable (e.g. solar); (8) Other.
- 472: Do any of the following reasons apply? (1) I prefer cooler temperatures; (2) I cannot afford to have the home as warm as I would like; (3) It is hard to keep the home warm because it is not well insulated; (4) None of these.
- 455: Does your home have a Building Energy Rating (BER) - a recently introduced scheme for rating the energy efficiency of your home? (1) Yes; (2) No; (3) Don't know.
- 5418: Moving on to education, which of the following best describes the level of education of the chief income earner? (1) No formal education; (2) Primary; (3) Secondary to Intermediate Cert Junior Cert level; (4) Secondary to Leaving Cert level; (5) Third level; (6) Refused.
- 4021: Can you state which of the following broad categories best represents the yearly household income BEFORE TAX? (1) <15 000 EUR; (2) 15 000-30 000 EUR; (3) 30 000-50 000 EUR; (4) 50 000-75 000 EUR; (5) >75 000 EUR; (6) Refused.

ANNEX D: Personae description

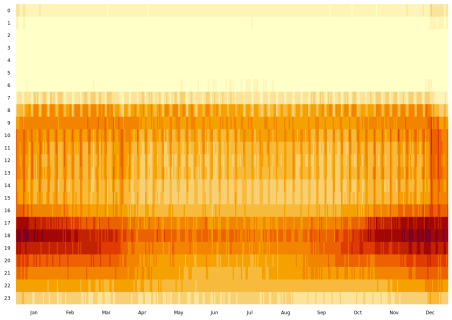
The reports of results of all the experiments carried out in this document can be found in the Zenodo repository entitled '*Analysis Reports on "Time Series from Smart Meters" Dataset*' (https://zenodo.org/record/6344956). Here, the 'top 40 clusters' obtained from the five selected datasets and their persona description can be found (see Section 6.2.1).



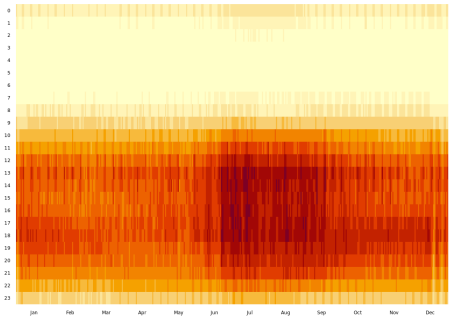
#1
Inhabitants working outside home. The weekly schedule covers leaving home from 8:00 am to 10:00 am and getting back at 17:00 pm. Weekend schedule is different though: staying at home on Saturday mornings and staying up late; on Sundays, resting at home in the mornings all year around but going out on Sunday afternoons, except in winter time. In winter, there is more time spent at home during evenings and the consumption is quite high: 7.5 ± 3 MWh/yr, possibly because of (additional) electric heating. Holiday season is spent at home.



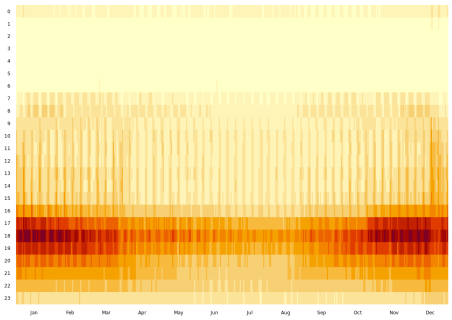
#2
This profile has a very strong seasonal component with highest consumption concentrated in the colder months of winter and spring. There is no apparent weekly pattern. The electric heating is on throughout most of the daytime (not between midnight and 8:00 am), with broad peaks around lunch and dinner time. Inhabitants are at home most days of the week (they do not work or work from home) and prepare meals on an electric cooker. Despite the electric heater, their overall electricity consumption is not very high: 5 ± 2 MWh/yr. This might indicate that they live in well insulated homes, or heat only the rooms used during the day, and/or that the family unit is small (few family members). They go to bed fairly early (22:00 -23:00 pm). They do not seem to go on summer holidays, but this is unclear as consumption is low in summer anyway. Probably a profile of elderly or retired people.



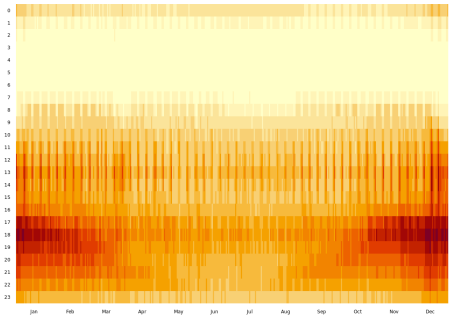
#3
There is someone at home all day during weekdays, but higher consumption/occupation at weekends and during Christmas holidays. Very strong early evening peak, no lunch time peak. Early to bed. Fairly high consumption (avg. 6MWh/yr). This is probably a family with young children, one homemaker who does not cook lunch for him/herself but does cook dinner for the other family members.



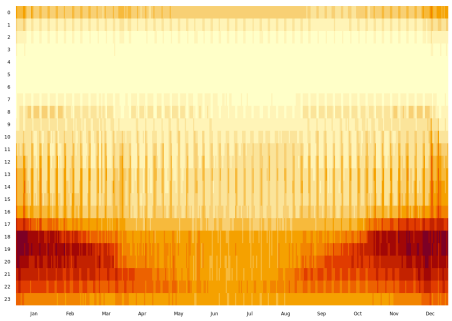
#4
Consumption concentrated broadly during daytime, from mid-morning until mid-evening. No clear weekday/weekend pattern, but strong increase during hotter months (likely due to air conditioning use). The overall electricity consumption is not high (4MWh/yr average but heavily skewed towards higher values). Likely not a private home but a very small business.



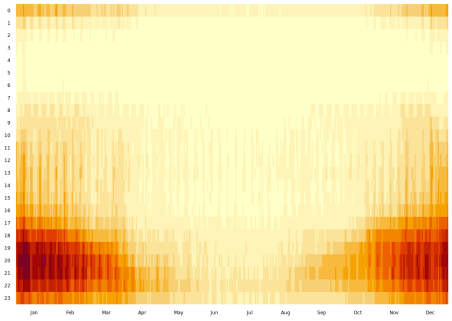
#5
All consumption is strongly concentrated around the early evening peak (18:00 pm). Family members are out of the house during the day on weekdays and gather for dinner. They go to bed quite early. They spend weekends in or around home, especially in winter. They follow the same pattern in summer holidays.



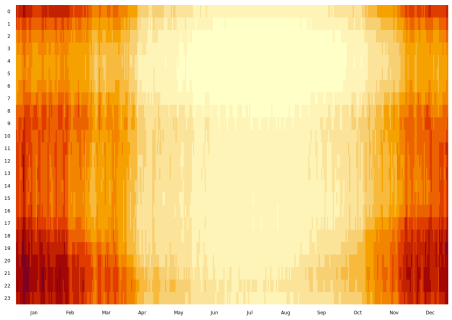
#6
Possibly there is someone at home all day during weekdays, but markedly higher consumption/ occupation at weekends (especially lunchtime) and during Christmas holidays. Maybe one person boils the kettle or makes toast for a light lunch, but the main meal is cooked at dinner time for the whole family. They do not appear to go away for the summer. Very high average consumption of 9MWh/yr, so there might be some electric heating (and/or DHW).



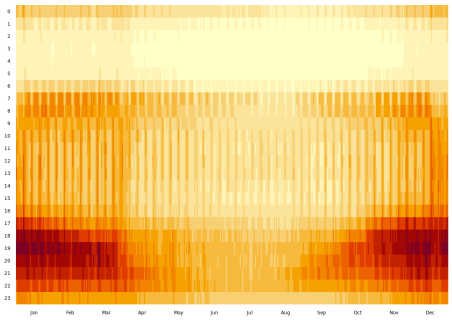
#7
Consumption sharply concentrated on early evening peak, especially during the darker/colder months. Work away from home. Do not come home to lunch. Cook dinner early in the evening. Weekends at home. Stay up fairly late. No clear holiday effect. High consumption: 8MWh/yr, broadly spread.



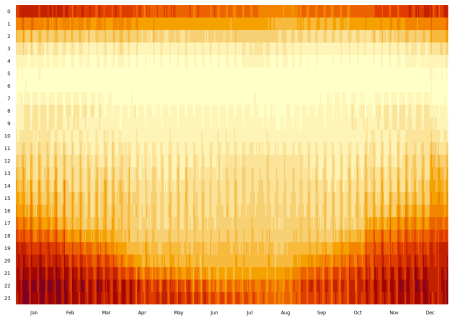
#8
Probably a residential home, people leaving in the morning with only a cold breakfast. Some consumption during the day, especially in the afternoons—there are probably children. Strong daylight/temperature correlation. Stay at home during the weekends. Go to bed at 01:00 am.



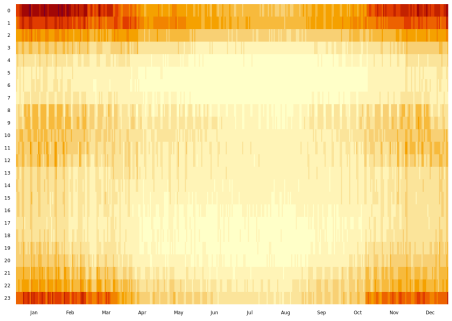
#9
Electric heating, especially strong after 16:00 pm during the winter; turned down during working days.



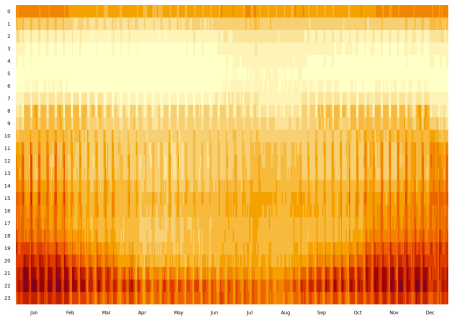
#10
Working away from home. Arrive at home at 17:00 pm. Prepare and have dinner between 18:00-22:00 pm. Then, staying up until 01:00 am. Saturday nights going out. Holidays at home.



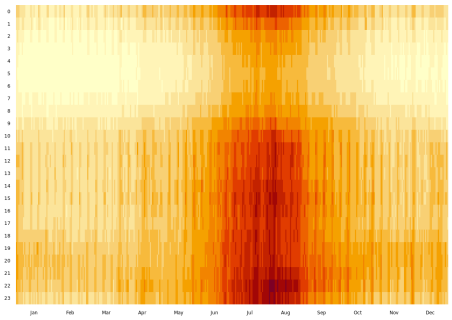
#11
Probably a residential home, working away from home. Get up between 6:00 and 7:00 am, leave for work around 9:00 am. Back home between 16:00 and 18:00 pm. Significant temperature correlation. Summer holidays and Christmas season, at home. Mostly at home during weekends.



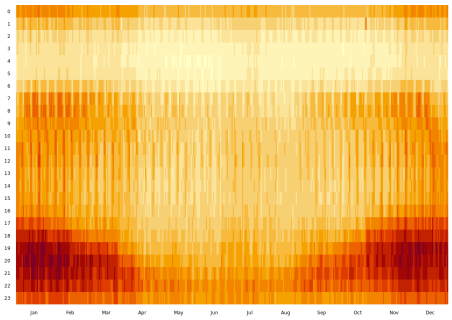
#12
Probably use of an electric water boiler—most likely supported by solar thermal. Heating on from 23:00 pm to 3:00 am, with some additional heating during the day, after people shower in the morning. Staying up late on weekends.



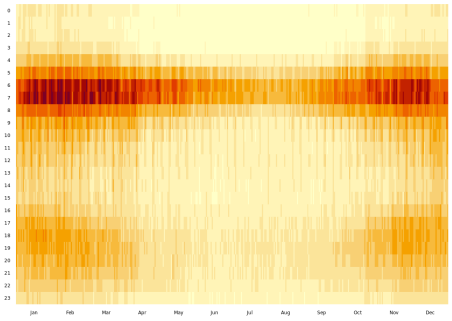
#13
Probably a residential home. Low sleep need. Stay up very late during the summer and weekends. Lunch at home. Significant energy demand in the evenings - there might be a big media hobby. Holidays spent at home. During summer vacation, probably teenagers get up around 10:00 am and turn on all the electronics. Weekends at home with strong energy consumption all day.



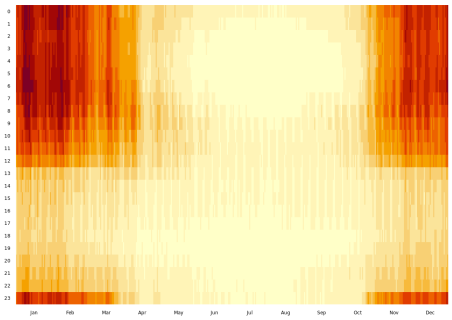
#14
Probably a spare home used mostly weekends and long summer vacation. Probably strong air conditioning.



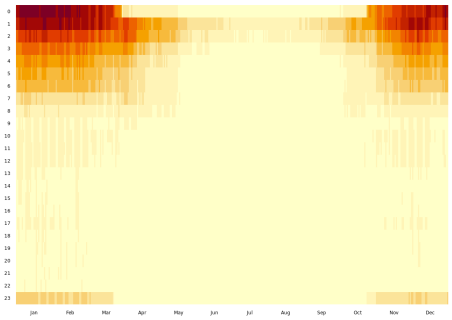
#15
Same pattern during the week and weekends. Consumption between 20:00 pm and 01:00 am. Holidays away.



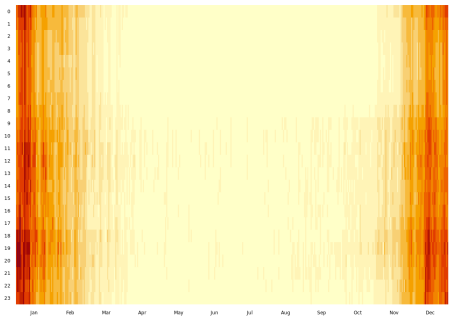
#16
Working away from home. Waking up at 16:00-17:00h. Biggest consumption between 6:00 am and 7:00 am. Then leaving home at 9:00 am. Coming back home at 17h. Holidays at home



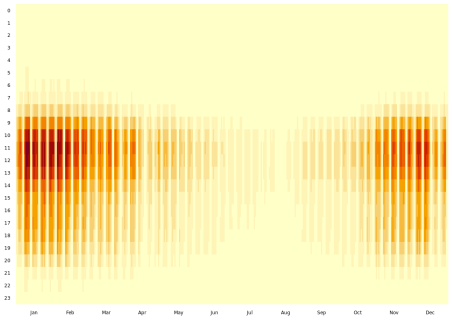
#17
Thermal energy storage. Used from 23:00 pm to 12:00 am. Mainly programmed off-peak accumulation heaters. High consumption (6 ± 2 MWh/yr). Holidays at home. Do not have a spare house.



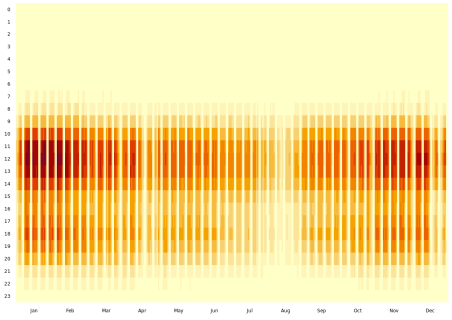
#18
Thermal energy storage. Used from 23:00 pm to 06:00 am. Mainly programmed off-peak accumulation heaters.



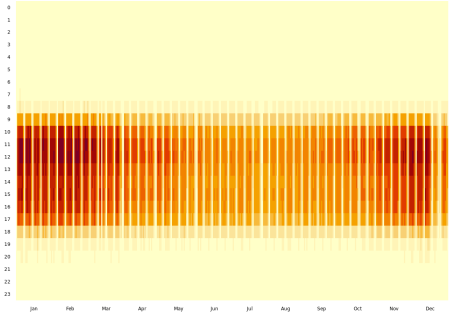
#19
Electric heating. Used in winter all day.



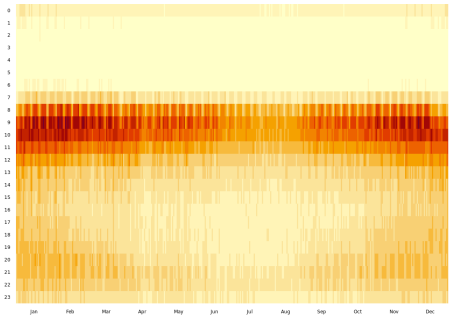
#20
Probably office apartments. Working between 9:00 am-15:00 pm and 17:00 pm-21:00 pm. Consumption is bigger between November and March, probably because they use heating in these months. During the weekends, holidays and summer is empty. Light driven



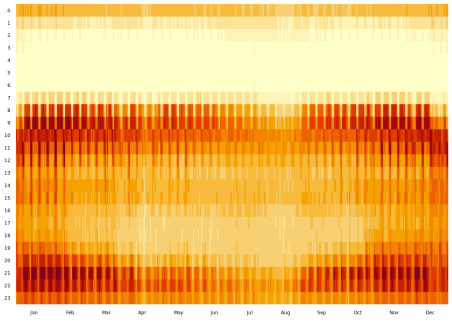
#21
Probably office apartments. Working between 9:00 am -15:00 pm and 17:00 pm - 21:00 pm. During the weekends, holidays and summer is empty.



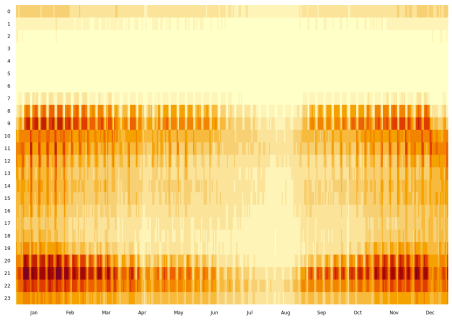
#22
Probably a small business. All year open. Very regular, 9:00 am -18:00 pm. Sundays are off. Very small consumption on Saturdays, peak consumption 11:00 am -12:00 pm. High consumption hours increase progressively when approaching winter, Jan-Dec, less during Christmas break. High consumption hours decrease from Jan to May. Less consumption: summer, Christmas, Eastern holiday. Triangular pattern of peak consumption hours, increasing towards winter, then decreasing.



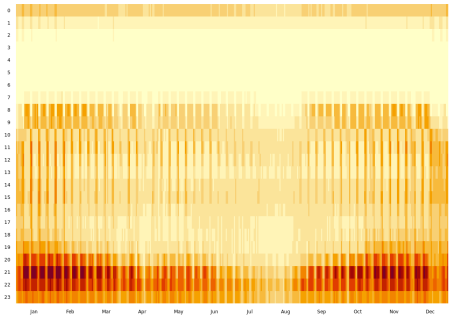
#23
Profile working away from home. Biggest peak in the early morning. Arrive at home at 20:00 pm. Dinner at home. Triangular shape on consumption hours. Less consumption on weekends but it is still there. Little consumption after 1:00 pm, then a bit more in the afternoons. Almost no consumption during the summer.



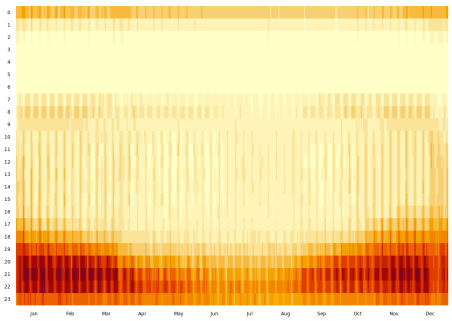
#24
Probably someone is at home all day. Breakfast and dinner at home. Electric kitchen. Strong consumption mornings (8:00-9:00 am) and nights (20:00-22:59 pm), less on weekends. Weekends stronger in the mornings, starting at 10 am, not during the evening (from 10:00 am to 15:00 pm). Consumption decreases during the summers. Break rhythm in August and holidays at home.



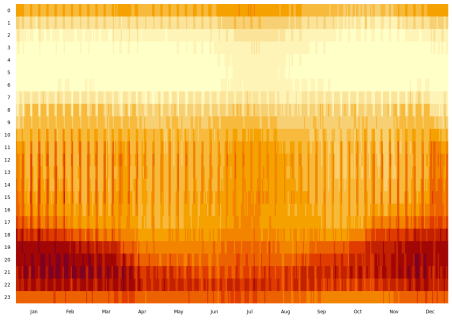
#25
It is probably someone at home all day during weekdays. Lunch and dinner at home. Going out on Saturday nights. They do not spend much time at home during weekends, especially in summer. Break rhythm in summer.



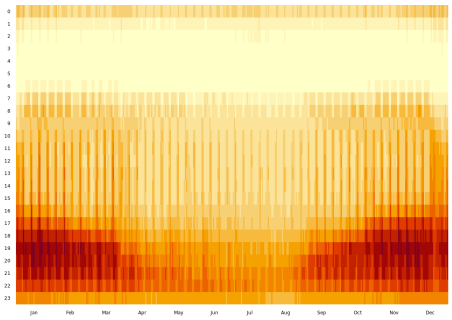
#26
Probably working away from home. More consumption in the evenings during weekdays, more consumption in the morning until after lunch on the weekends. The pattern seems like a home, no breakfast, but dinner at home. On the weekends, breakfast/lunch at home, but not dinner at home Saturdays and Fridays. Sundays activity all day, resting at home. Breaks rhythm for summer holidays, Easter and Christmas (more away from home). Less activity during summer.



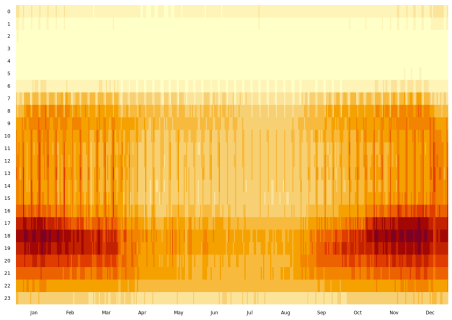
#27
Working away from home during weekdays. Dining on weeknights. Going out on weekend nights, then staying up late at home probably after the party. On Sundays, resting at home. Breaks rhythm for summer holidays, Easter and Christmas (more away from home). Probably it is young people with active social/night-life. No electric heating. Have no spare house for weekends.



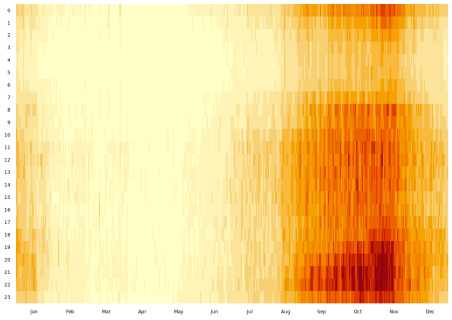
#28
Probably someone is working from home. Strong consumption during the evenings and weekends all day. During the summer and Christmas holiday, consumption goes up (probably because there are more people at home), and shifts up on weekdays and down at weekends, so that there is almost no difference between weekdays and weekends. During weekends, they sleep more and go out at night. Do not have a spare weekend house. Summer holidays are spent at home. No electric heating.



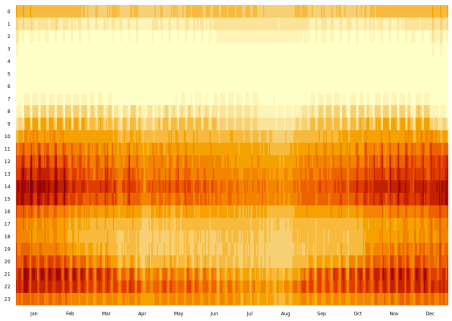
#29
Probably working away from home. Leaving home around 9.00 am and coming back at 17:00 pm. No lunch at home. They cook and use home appliances between 18:00 to 22:00 pm - same pattern some/most weekends. Weekends spend at home. Holidays mostly at home. Average consumption. Probably no electric heating.



#30
Probably a residential home with a very strong week-weekend pattern. No electric kitchen and no occupancy during the day (only in the afternoon from 17:00 pm). Energy consumption is strongly correlated with daylight (and does not seem to be with climate). The patterns change at weekends and energy consumption seems quite regular there. There are probably no electric heaters at home.

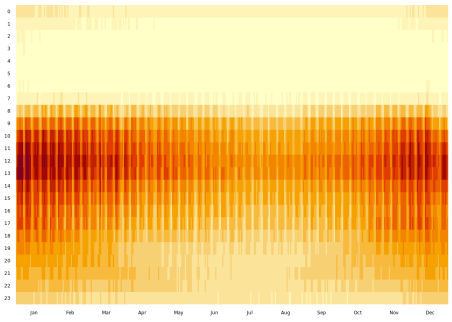


#31
Probably a spare home related to the harvest of apples or grapes. Very low energy consumption centred at Autumn period. There is clearly daylight related consumption, increasing at night and regular energy consumption during the day. No consumption late at night.



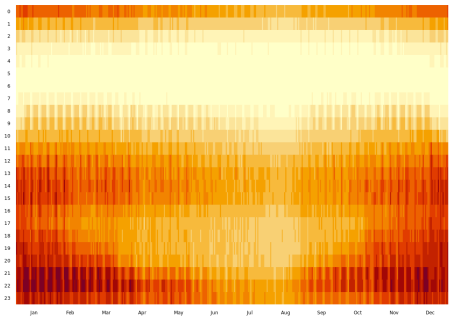
#32

Probably a residential home. Clear pattern week-weekend and less clear on holidays, so probably there is a spare house to go on holidays or there is a substantial change of behaviour between weekdays and weekends. During weekends, they spend more time at home or probably there are more people at home. There is an electric kitchen. Inhabitants have lunch and dinner at home. Moreover, the energy consumption is correlated with the climate so have also electric heaters. Another possibility at winter time is that inhabitants stay at home during the morning preparing lunch. At nights, there is a small correlation with daylight, but is mostly camouflaged by the dinner consumption.



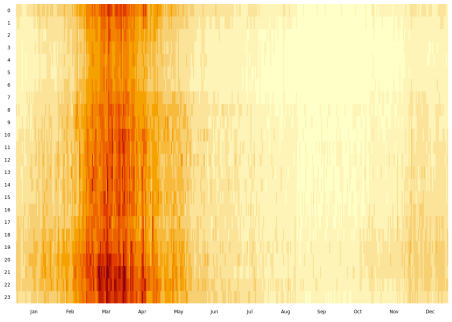
#33

Probably a public authority building such as a library, a community centre. Only consumption during weekdays and very regular, starting at 9:00 am and finishing off at 17:00 pm. Small energy consumption starting at 8:00 am probably for cleaning and after 17:00 (after school activities). There are electric heaters as the consumption has a climate correlation.



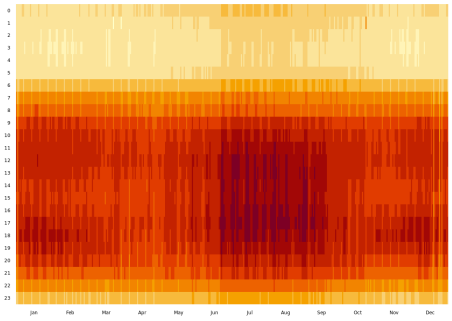
#34

Probably a residential home. Clear pattern week-weekend and holidays, so inhabitants change substantially the behaviour between weekdays and weekends. There is a strong daylight related pattern but not a relevant weather one. They seem to be at home for lunch and dinner but it seems like they do not have an electric kitchen. There is siesta / going out time in summer. The reduction in summer nights could be just for having cold dinner plus going out.

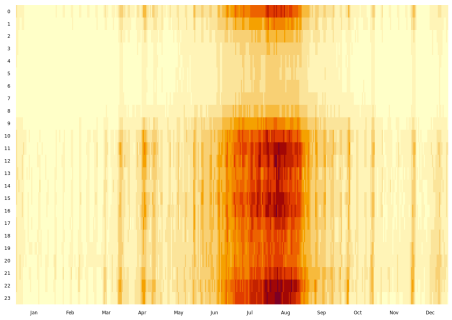


#35

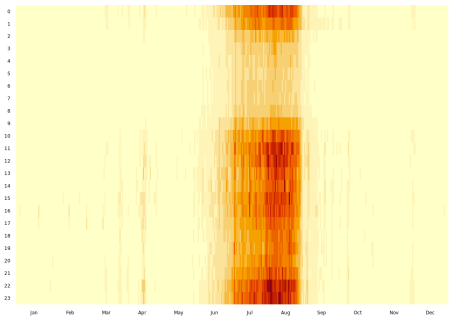
Probably a spare home used at the end of winter and early spring, maybe related to winter sports. The pattern seems quite irregular with consumptions all over the day (including late at night) which could suggest the use of electric heaters. There is a clear increase of energy consumption at night so most probably they have dinner at home but the rest of the day the behaviour is more irregular. The house is less used the rest of the year (or their energy consumption is far lower than during the winter-early spring period).



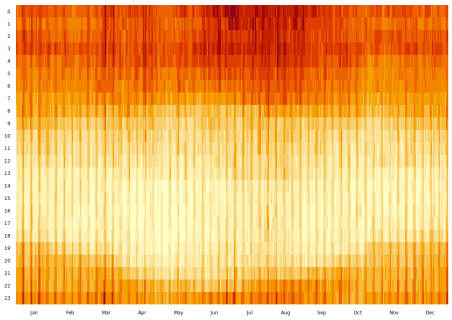
#36
Probably a small supermarket: opens regularly since 9:00 am to 21:00 pm. Energy consumption from 6:00 am for resupply and until 22:00 pm to make the checkouts. Small difference between weekends and weekdays and probably only closed on Sundays and long weekends at Christmas season. Slightly more energy consumption in summer which suggests the increased use of air conditioning.



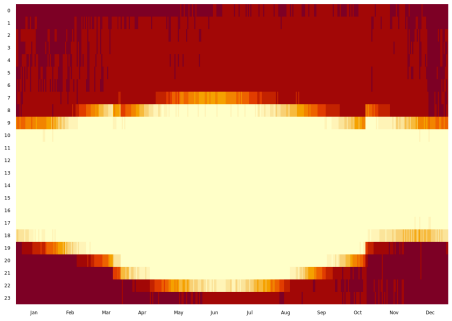
#37
Probably a spare house, occupied almost all weekends year long (specially on long weekends) and holidays. The energy consumption is regular during the day with peaks at breakfast, meals and dinners. No visible differences between weekdays and weekends at summer time.



#38
Probably a spare house, occupied at summer time and long weekends (mostly in Easter week and other summer holidays). Regular consumption during the day with peaks at breakfast, meal and dinners. No visible differences between weekdays and weekends during the summer period. The probability to start this pattern increases in late June (following the school holidays) and peaks in August. The probability stops abruptly in September.



#39
Lights from a neighbourhood community, which includes other energy consumption during the day but is not as regular as lights at nights.



#40
Street lights from a municipality.

ANNEX E: Taxonomy

To analyse in detail the forty electricity consumption profiles obtained, a taxonomy based on the similarity of the electricity consumption patterns of each profile has been developed. The taxonomy is represented using a dendrogram in Fig. E1. The number at the end of each branch refers to the personae descriptions listed in **Annex D**.

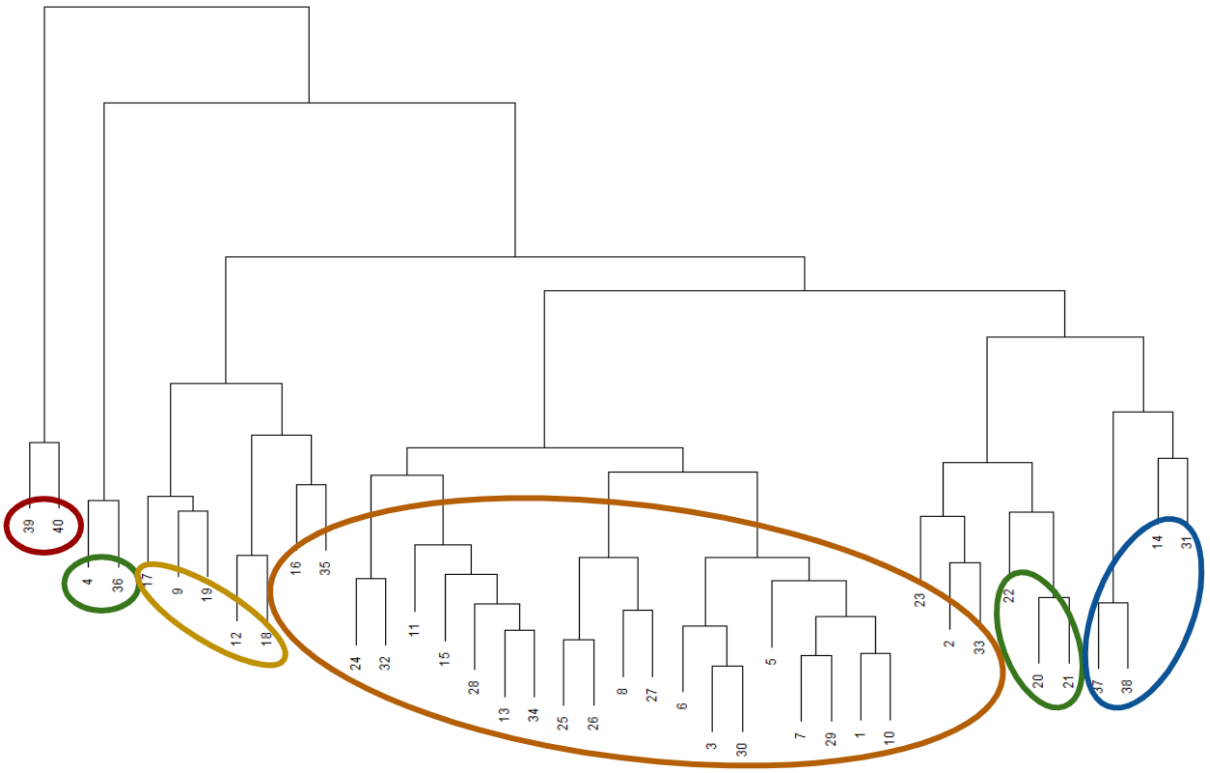


Figure E1: Dendrogram highlighting the most relevant groups.

As indicated by the coloured ellipses, five well-differentiated groups of consumers can be identified: on the one hand, the largest group, in orange, contains almost three quarters of the load profiles analysed, and is formed by **main households**. On the other hand, it can be distinguished a group formed by **secondary houses** (those that are occupied only in summer or holidays), in blue; a group where consumption is more associated with **equipment** (such as energy storage or electric heating), in yellow; a group of **offices and small businesses**, in green; and finally, a tiny group with consumption associated to **street lighting**, in red. Table E1 shows the distribution of time series by types of consumer:

Type of consumer	No. of sites	Percentage
Main houses	18 789	74%
Equipment	3 356	13%
Offices and businesses	1 688	7%
Secondary houses	1 235	5%
Street lighting	310	1%

Table E1: Distribution of time series by consumer types.

By zooming in on the largest group, i.e. main houses, a subdivision on the behaviour can be performed (see Fig. E2). A large subgroup of households, accounting for 30% of the total, in which behaviour **changes during weekends**, in green, can be found. Then, there exist households that spend **all day at home**, in purple; households that make more electrical consumption at midday, probably due to the use of **electric kitchens**, in yellow; and households where there are more pronounced consumptions around **all meal times**, in red.

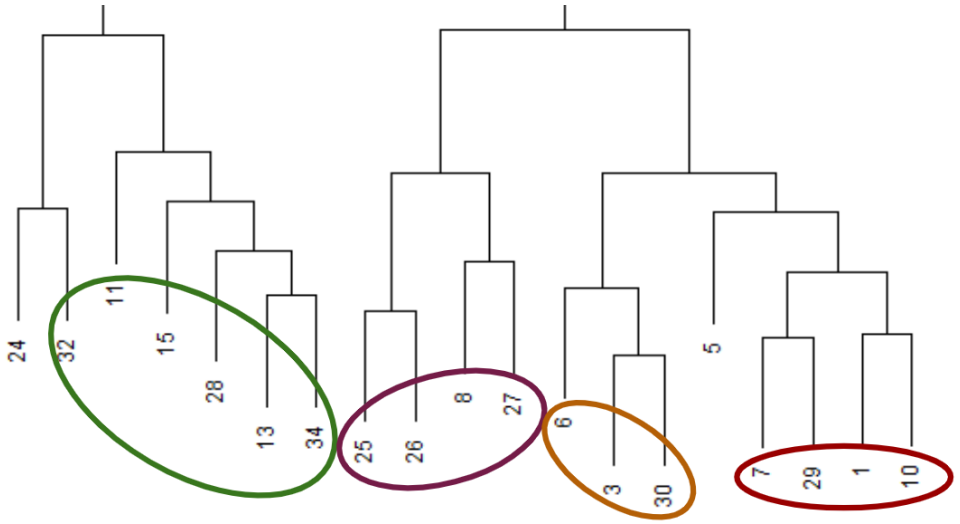


Figure E2: Zoom to the dendrogram at the 'main houses' group, highlighting the most relevant subgroups.

Table E2 shows the distribution of time series within this subgroup by types of consumer:

Type of consumer	No. of sites	Percentage
Changes during weekend	7 700	30%
All meal times	4 175	16%
All day at home	3 605	14%
Electric kitchens	3 309	13%

Table E2: Distribution of time series by consumer types within the 'main houses' group.