The BRIGHT project is co-founded by the EU's Horizon 2020 innovation programme under grant agreement No 957816

Boosting DR through increased communIty-level consumer engaGement by combining Data-driven and blockcHain technology Tools with social science approaches and multi-value service design

Deliverable
D4.4 DTs' model for customer's categorization – first version

Authors: Matthias Strobbe (imec), Gargya Gokhale (imec), Evelyn Heylen (Centrica), Bert Claessens (Centrica), Klemen Bregar (Comsensus), Andrej Campa (Comsensus), Giannis Kazdaridis (domX)

The project Boosting DR through increased communIty-level consumer engaGement by combining Data-driven and blockcHain technology Tools with social science approaches and multi-value service design (BRIGHT) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957816. The sole responsibility for the content of this publication lies with the authors. It does not necessarily reflect the opinion of the Innovation and Networks Executive Agency (INEA) or the European Commission (EC). INEA or the EC are not responsible for any use that may be made of the information contained therein.

Imprint

Title:	**DTs' model for customer's categorization – first version**
Contractual Date of Delivery to the EC:	31.01.2022
Actual Date of Delivery to the EC:	31.01.2022
Author(s):	Matthias Strobbe (imec), Gargya Gokhale (imec), Klemen Bregar (Comsensus), Andrej Campa (Comsensus), Giannis Kazdaridis (domX)
Participant(s):	IMEC, CEN, DOMX, COM
Project:	Boosting DR through increased communIty-level consumer engaGement by combining Data-driven and blockcHain technology Tools with social science approaches and multi-value service design (BRIGHT)
Work Package:	WP4 – Community and Customer Digital Twin Models
Task:	T4.4 – Digital Twins for customers clustering and segmentation
Confidentiality:	Public
Version:	1.0

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

API	Application Programming Interface
BRIGHT	Boosting DR through increased communIty-level consumer engaGement by combining Data-driven and blockcHain technology Tools with social science approaches and multi-value service design
DB	Database
DR	Demand-Response
DT	Digital Twin
GUI	Graphical User Interface
JWT	JSON Web Token
NILM	Non-intrusive Load Monitoring
REST	Representational State Transfer
std	Standard deviation
WP	Work Package

Table 1 List of Acronyms and Abbreviations

Executive Summary

The objective of T4.4 is to develop and evaluate clustering and segmentation algorithms for residential electric and thermal load profiles. These algorithms will allow to cluster flexible loads into different behavioural clusters and via energy disaggregation algorithms to get more insight in load patterns on building or household level. The resulting algorithms help to further improve the digital twins that are developed in T4.3 (individual flexible devices) and T4.5 (groups of flexible devices in communities) as well as the energy forecasting algorithms that are being developed in T4.2. These digital twins and algorithms serve as input for the development of consumer-centred flexibility services in WP5.

This deliverable presents the first results on clustering and segmentation:

- A literature study and proposed methodology to cluster electrical and thermal profiles from residential users that are part of the Belgian pilot. Different clustering techniques and distance metrics will be compared and user related information (from WP2 and WP3) will be used to further refine and interpret the obtained clusters. These clusters will then be used to improve the digital twins that are developed mainly in T4.5 (described in D4.5).
- A clustering algorithm for residential gas boilers, as part of the Greek pilot, where the main goal is to compare the developed adaptive control strategy with the legacy control scheme in terms of reduced gas consumption and improved experienced comfort for the end user.
- An ensemble of 4 segmentation algorithms for residential profiles to identify the consumption profiles of individual devices within these households. As such a better insight can be obtained on typical user activities and the potential flexibility from controllable devices within the homes such as heat pumps, boilers, whitegoods, etc. The results are also useful for other types of services, e.g., to detect anomalies in individual devices or in user activities/behavior. The latter could be interesting for assisted living scenarios.

1. Introduction

To efficiently valorize the flexibility of flexible loads via demand response programs, accurate models are needed for these controllable loads. Within BRIGHT we focus especially on residential flexible assets which are very dependent on specific user behavior. The clustering of flexible load patterns in different behavioural clusters allows to further refine the digital twins models that are developed in T4.3 and T4.5, and improves load and flexibility forecasting, as developed in T4.2. More accurate digital twins allow to improve the flexibility services developed in WP5 (e.g. change of offer based on typical energy consumption behavior) and thus increases the value of residential flexibility.

1.1. Purpose

This deliverable presents the first version of the clustering and segmentation algorithms that are developed or are under development to process electric and thermal load profiles in combination with available user data. In the second version of this deliverable (to be released in M30) further optimized algorithms and the validation on larger (pilot) datasets will be detailed. The results are valuable to different actors involved in the design, implementation and exploitation of DR products and services (e.g. energy services providers, aggregators, grid operators, energy suppliers, energy communities, etc.).

1.2. Relation to Other Activities

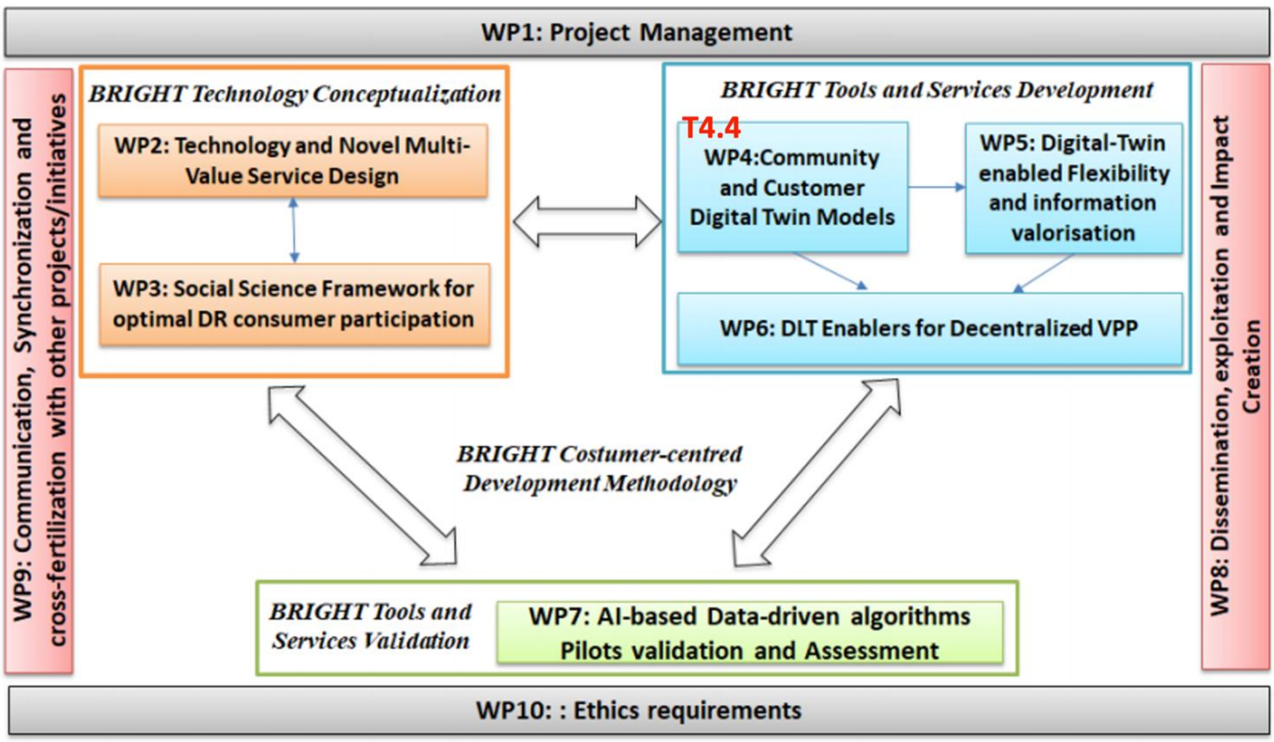


Figure 1. BRIGHT WP interaction overview

This deliverable is the output of Task 4.4 of work package 4. It uses the outputs from WP2 and WP3 on requirements, use cases, and drivers and barriers for consumer engagement as input for the development of the clustering algorithms. The results are then used to improve the results from the other WP4 tasks. The combination of clustering algorithms, digital twin models and demand response services will be tested and validated in the BRIGHT pilots as part of WP7.

1.3. Structure of the Document

This deliverable is organized as follows:

- Section 2 presents the methodology that will be used to cluster residential heat and electricity consumption profiles for the Belgian pilot.
- Section 3 presents the developed clustering algorithms for residential gas consumption, to compare energy savings between the smart control and legacy control schemes for the Greek pilot.
- Section 4 presents results on energy disaggregation algorithms for households, which allow to get a better insight in the specific profile of a residential user and the potential flexibility they could offer.
- Section 5 presents the conclusions and the planned next steps, which will be reported in the second deliverable for this task.

2. Clustering of Residential Electricity and Heat Consumption Profiles

2.1. Introduction

A significant amount of data is being generated in residential households due to improvements in metering and data collection systems. This data provides information related to the consumption patterns and flexibility potential of different households, which can be leveraged for improving the performance of various models and control algorithms being developed within the BRIGHT Project (WP4, WP5). Processing and analyzing large amounts of individual profiles requires a lot of computation and storage resources, as well as extensive security mechanisms to secure this privacy sensitive information, and thus might not be very scalable. Via clustering, the most relevant insights can be retained in a more scalable and privacy-friendly way.

2.2. Literature Review

Previously, such categorizations were carried out based on socio-demographic information of individual households such as number of occupants, age, gender etc. While this provides an intuitive method for clustering, such methods are inefficient and can lead to inaccurate categorizations. Following advances in machine learning, data-driven clustering methods are becoming increasingly popular for such clustering tasks in the residential energy consumption domain [1] [3] . In [1] , a K-means based clustering method is studied in the context of residential electric water heaters for demand response. The authors discuss the clustering and control strategy, where the available data is first categorized into 2 groups and then a dynamic programming-based control strategy is studied. This work uses a simple K-means clustering technique to identify clients that have higher consumption during peak power periods. The number of clusters is first determined using the Silhouette index. The authors argue that developing control strategies for entire groups is beneficial and leads to an improvement in performance. Similar to this study, [2] focuses on residential buildings and presents a bottom—up analysis of energy consumption modeling in residential buildings. Toeffectively model the energy consumption, the authors first identify occupant behavior patterns which are then used for efficiently modeling the energy consumption. A k-modes clustering algorithm is used to accommodate for categorical input data and the number of clusters is determined using the Akaike Information Criterion. Additionally, a neural network is employed to map occupant's personal information to the clustered behaviors. Unlike works presented in [1] & [3] , this work forms clusters based on occupant's socio-demographic data and do not directly incorporate energy related data points. This makes the presented approach highly sensitive to the accuracy of clustering and is a limitation of this method. While works [1] & [2] focus on individual assets like water heaters, the study presented in [3] focuses on residential consumers that are part of a district heating network. The authors of this work use the k-means clustering technique to identify different patterns in time series data corresponding to daily load profiles. The authors identify different patterns in the heat consumption profiles and discuss the feasibility of clustering-based load forecasting. The authors argue that households follow typical consumption patterns that have high chances of repeating themselves and this can be used to effectively forecast energy consumptions. These works demonstrate the role of clustering in heat and electricity consumption data and provide methodologies for these tasks.

We aim to leverage these works, to identify clusters that consider socio-demographic data obtained from WP2 and WP3, along with data obtained from the smart metering infrastructure installed at the pilot sites.

2.3. Proposed Methodology

The electricity and heat consumption data for the Belgian pilot is collected for individual households and represents a time-series dataset as shown in Figure 2 and Figure 3. Heat is provided via a district heating network which is supplied with heat from different sources: most heat is waste heat from a nearby soap factory. If necessary, extra heat is supplied via heat pumps and/or a gas boiler.

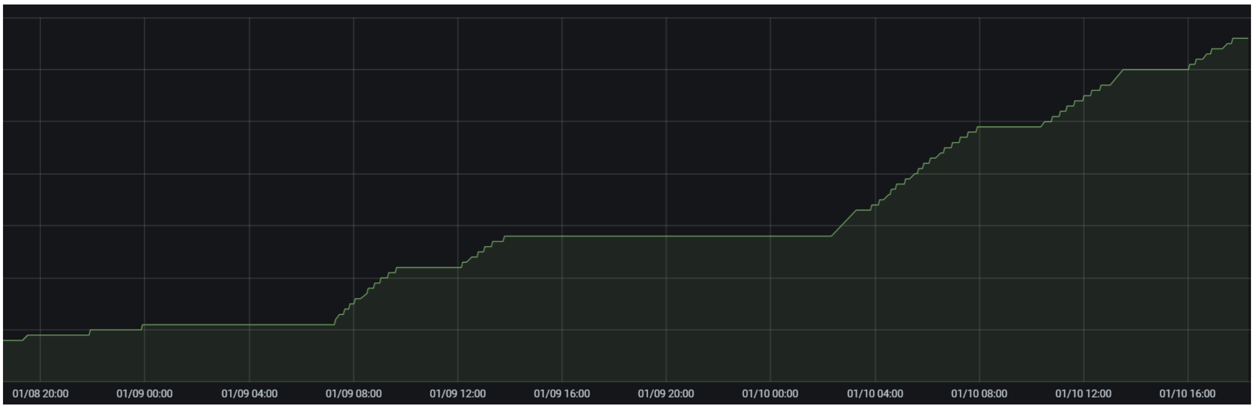


Figure 2: Cumulative heat consumption profile for a household

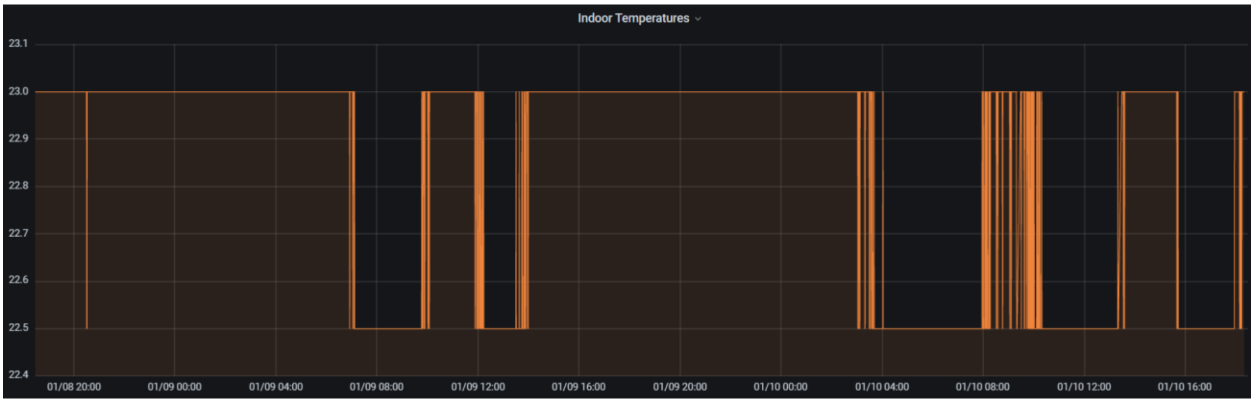


Figure 3: Indoor room temperature profile for a household

Such profiles are produced for each household for all days of the year, leading to a large dataset of different profiles. To effectively use this data, it is important to first identify similarities in different consumption profiles and group them into different categories. As presented in [3] , different clustering techniques can be utilized for this task. These techniques can be categorized into 3 main categories – partitional, density-based and hierarchical. Further, different distance metrics such as Euclidean distance, shape-based distance can be used with these clustering techniques. For selecting the appropriate technique and distance metric, authors in [3] recommend using a pool of clustering methods that produce different clusters with higher spreads. With this approach, the authors argue that a suitable clustering technique can then be chosen based on domain knowledge or other external measures. Following these recommendations, we will implement a cluster pool by focusing on three main clustering techniques i.e., K-means, Agglomerative and Density Peaks. Additionally, different distance metrics such as Euclidean distance and shape-based distance will be analyzed with the K-means algorithm. Following a preliminary clustering, outputs from WP2, WP3 will be used to interpret and refine these clusters.

3. Clustering of Residential Gas Consumption Profiles

Residential heating systems with boilers distribute heat with hot water, which is circulated through radiators or other such devices placed inside the house. A room thermostat which is directly connected to the boiler continuously measures the indoor room temperature and depending on the desired room temperature (setpoint) of the user, it controls the boiler activation. *Legacy* heating systems set a fixed heating temperature for the water circulated through the radiators, typically within the range of 65-80°C, irrespective of the indoor and outdoor conditions as well as the desired room temperature set by the user. This behaviour may lead to inefficient heating patterns and costly energy expenditure.

domX has developed a smart monitoring and control system which involves the installation of a heating controller that acts as a bridge between the thermostat and the boiler. The domX controller runs a control loop algorithm to control the boiler activation patterns and adapt the actual temperature of the water circulated through the installed radiators based on indoor, setpoint and outdoor temperatures. The **adaptive** mechanism aims towards improving the boiler efficiency and the perceived user comfort whilst reducing the energy consumption and costs.

In order to compare the energy savings between the two modes of operation, legacy and adaptive, and initiate the creation of gas consumption profiles, a multi-purpose clustering tool has been developed which is enabled by machine learning techniques and currently supports the automatic extraction of comparable residential heating scenarios and the production of comprehensive human-readable gas consumption evaluation reports.

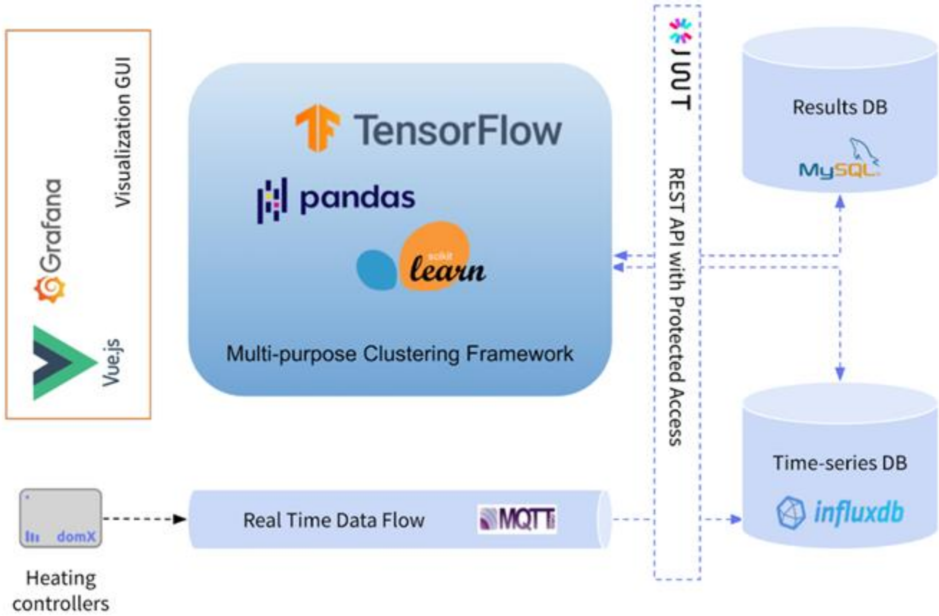


Figure 4. High-level architecture of multi-purpose clustering framework.

The high-level architecture of the clustering framework is shown in Figure 4. Real-time data collected from the heating controllers are integrated using the MQTT message broker and stored in the Influx time-series DB. Automatic data handling and model training pipelines have been developed using the Python libraries Pandas, SciKit Learn and TensorFlow. The clustering results are stored in the MySQL DB and exposed over a REST API and can only be accessed using JWT tokens to guarantee secure data exchange with other services. A custom web-based graphical user interface has been developed by combining the Vue.js and Grafana frameworks for visualizing the outputs of the multi-purpose clustering tool.

3.1. Data acquisition and pre-processing

The raw data recorded for every home exist in buckets that contain various time dependent metrics related to (a) indoor and outdoor (outside the building) metrics, e.g., temperature and humidity, (b) user preferences, e.g., desired setpoint temperature and heating balance and (c) boiler activity, e.g., water temperature and boiler modulation level. These values have been captured for several months. The developed data analysis process works on the basis of extracting numerical features from the continuous time series data in order to summarize daily heating scenarios as well as daily consumption. As such, each one data sample in the pre-processed set consists of the extracted daily features which characterize the corresponding day from the recorded period.

Since the raw data exist in various sampling frequencies, some simple pre-processing steps were performed to polish the data, i.e., (a) resampling all measurements to 1 second intervals, (b) removing erroneous values using rule-based outlier detection (e.g., known extreme temperature values) and (c) handling missing values by linear interpolation or filling.

3.2. Feature extraction

The set of extracted daily features are related to the conditions where heating was requested as well as its duration:

- **mean out temp**: the mean value of all outdoor temperature data points
- **std out temp**: the standard deviation of all outdoor temperature data points
- **mean target temp**: the mean value of all target room temperature data points
- **heating demand**: calculated as the hours of day where heating was active

These features were extracted and given as input to build the clustering model in order to group together days where both the outside conditions and the demand for heating were similar. As such, each day is represented as a vector in a 4-dimensional feature space. In addition, the following features are specific to the heating system characteristics and are calculated in order to display them in the clustering reports (those are not relevant to the clustering model):

- **boiler activations**: the number of boiler activations, from idle to heating mode
- **mean boiler temp**: the mean actual temperature of the water circulated between the boiler and the radiators
- **mean boiler max temp**: the mean requested temperature of the water circulated between the boiler and the radiators
- **bypass**: this is the operation mode, legacy or adaptive
- **comfort**: a metric which indicates the amount of time that the actual indoor temperature is within +-0.5 degrees of the setpoint temperature, presented as a percentage (%) over the corresponding 24h period.

3.3. Clustering methodology

Hierarchical clustering is a general family of clustering algorithms which develops nested clusters by merging or splitting them successively. Agglomerative Clustering performs a hierarchical clustering using a bottom-up approach: each observation starts in its own cluster, and clusters are successively merged together. The linkage criterion determines the metric which is used for the merge strategy. Specifically, maximum or complete linkage minimizes the maximum vector distance between observations of pairs of clusters and was adopted in order to create clusters of comparable heating scenarios (days). This can be done by defining the maximum allowed feature vector distance which can exist between any two days within a particular cluster. The Manhattan distance was chosen in order to calculate the distance between feature vectors. The distance is simply calculated as the sum of absolute differences of all features and is easy to interpret and tune.

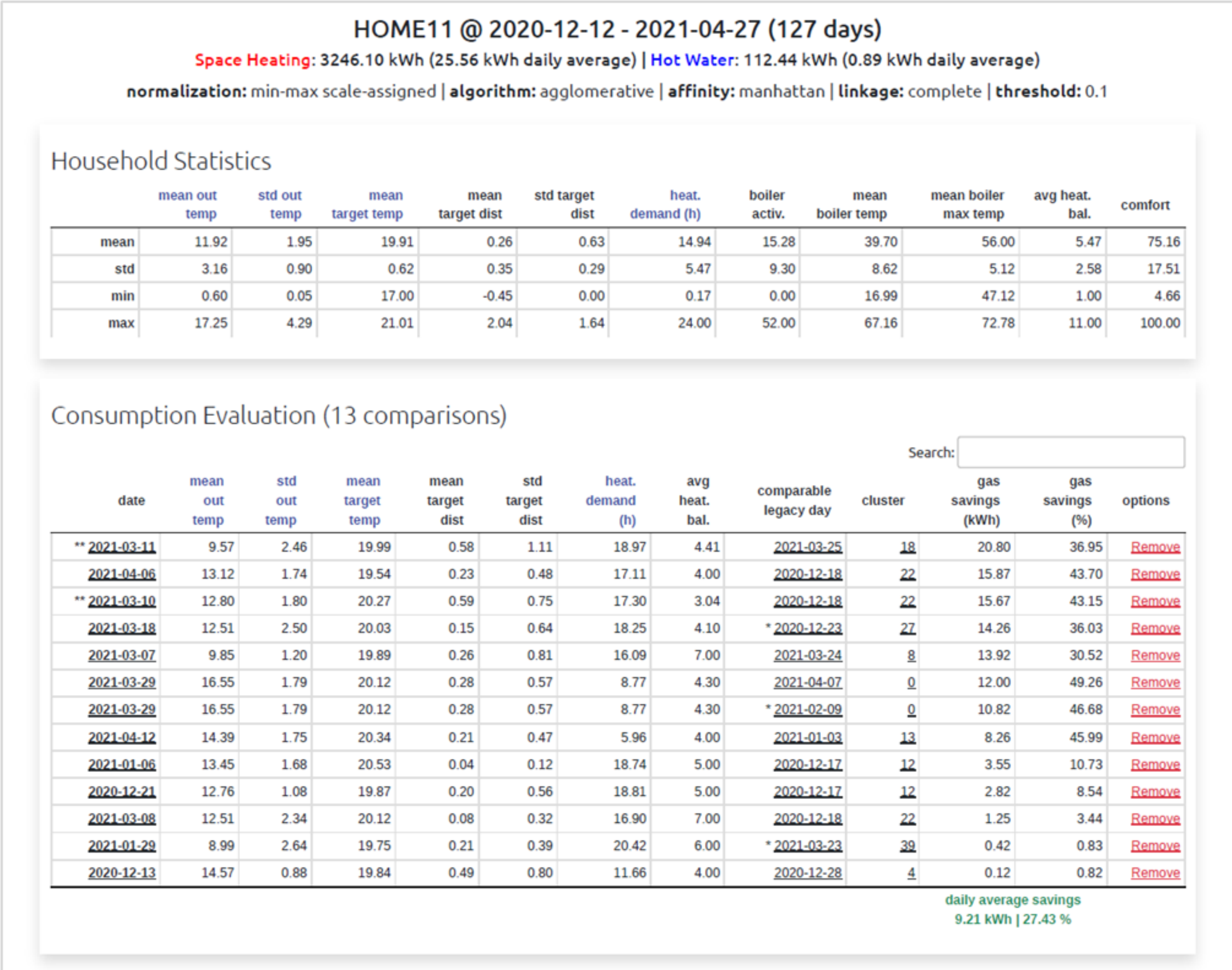
After some initial experimentation, the mean outdoor temperature and the heating demand duration were empirically determined as the two most important factors which affect the amount of consumed gas in a particular scenario. In order to reflect the higher impact of the mean outdoor temperature and the heating demand when producing the clusters, a feature normalization step was designed such that the calculated distance of normalized feature vectors depend heavily on the most prominent features. This was achieved by producing a standard min-max normalization scheme while letting the min and max values of *std out temp* be defined by the min and max values of *mean out temp*.

The clustering outcome depends on selecting a distance threshold for which the linkage criterion stands true for all clusters. Since the clustering is currently performed per household, the optimal threshold value depends on the specific household and its feature space which is unique. By altering the distance threshold, the user can obtain clusters of varying cohesion. A lower threshold creates smaller and tighter clusters where the features match closely, while a higher threshold creates bigger groups where the features are more loosely related.

3.4. Clustering outcomes

After feeding the clustering framework with the data from several domX households, the clustered heating scenarios are extracted for each home and the corresponding human-readable reports are produced. The clustering outcomes can then be used to monitor various aspects of the operation of the heating system between comparable heating scenarios. Specifically, those have already been used to measure the energy savings of the domX adaptive operational mode over the energy requirements of the legacy operation for each household.

As an example, the first page of the produced consumption evaluation report for domX client *Home 11* is shown in Figure 5. The home name and the chronological period under investigation are shown in the title. Bellow, the total and average daily gas consumed for space heating and for domestic hot water are shown. In another subtitle, information about the clustering parameters that produced the report are shown. The user can be informed about some basic statistical metrics for each calculated feature for the corresponding chronological period in the *Household Statistics* section.



HOME11 @ 2020-12-12 - 2021-04-27 (127 days)

Space Heating: 3246.10 kWh (25.56 kWh daily average) | Hot Water: 112.44 kWh (0.89 kWh daily average)

normalization: min-max scale-assigned | **algorithm:** agglomerative | **affinity:** manhattan | **linkage:** complete | **threshold:** 0.1

Household Statistics

	mean out temp	std out temp	mean target temp	mean target dist	std target dist	heat. demand (h)	boiler activ.	mean boiler temp	mean boiler max temp	avg heat. bal.	comfort
mean	11.92	1.95	19.91	0.26	0.63	14.94	15.28	39.70	56.00	5.47	75.16
std	3.16	0.90	0.62	0.35	0.29	5.47	9.30	8.62	5.12	2.58	17.51
min	0.60	0.05	17.00	-0.45	0.00	0.17	0.00	16.99	47.12	1.00	4.66
max	17.25	4.29	21.01	2.04	1.64	24.00	52.00	67.16	72.78	11.00	100.00

Consumption Evaluation (13 comparisons)

Search:

date	mean out temp	std out temp	mean target temp	mean target dist	std target dist	heat. demand (h)	avg heat. bal.	comparable legacy day	cluster	gas savings (kWh)	gas savings (%)	options
** 2021-03-11	9.57	2.46	19.99	0.58	1.11	18.97	4.41	2021-03-25	18	20.80	36.95	Remove
2021-04-06	13.12	1.74	19.54	0.23	0.48	17.11	4.00	2020-12-18	22	15.87	43.70	Remove
** 2021-03-10	12.80	1.80	20.27	0.59	0.75	17.30	3.04	2020-12-18	22	15.67	43.15	Remove
2021-03-18	12.51	2.50	20.03	0.15	0.64	18.25	4.10	* 2020-12-23	27	14.26	36.03	Remove
2021-03-07	9.85	1.20	19.89	0.26	0.81	16.09	7.00	2021-03-24	8	13.92	30.52	Remove
2021-03-29	16.55	1.79	20.12	0.28	0.57	8.77	4.30	2021-04-07	0	12.00	49.26	Remove
2021-03-29	16.55	1.79	20.12	0.28	0.57	8.77	4.30	* 2021-02-09	0	10.82	46.68	Remove
2021-04-12	14.39	1.75	20.34	0.21	0.47	5.96	4.00	2021-01-03	13	8.26	45.99	Remove
2021-01-06	13.45	1.68	20.53	0.04	0.12	18.74	5.00	2020-12-17	12	3.55	10.73	Remove
2020-12-21	12.76	1.08	19.87	0.20	0.56	18.81	5.00	2020-12-17	12	2.82	8.54	Remove
2021-03-08	12.51	2.34	20.12	0.08	0.32	16.90	7.00	2020-12-18	22	1.25	3.44	Remove
2021-01-29	8.99	2.64	19.75	0.21	0.39	20.42	6.00	* 2021-03-23	39	0.42	0.83	Remove
2020-12-13	14.57	0.88	19.84	0.49	0.80	11.66	4.00	2020-12-28	4	0.12	0.82	Remove

daily average savings
9.21 kWh | 27.43 %

Figure 5. Example consumption evaluation report generated using the multi-purpose clustering framework.

The report results are shown in the *Consumption Evaluation* section (in parentheses, the number of comparisons that make up the final report is given). The report is presented as a large table where each row represents a single comparison between two similar days (heating scenarios) which were grouped into the same cluster: one where the domX adaptive mode was operating and one where the legacy operation was active instead.

The datetimes of a comparison pair are given in the 1st and 9th column and upon clicking they lead to the Grafana visualization dashboard for further exploration of the data on the corresponding dates. The features shown in each row are the calculated features of the adaptive day and the features relevant to the clustering algorithm are colorized purple. The clusters where similar days have been grouped together are also clickable and lead to the display of the corresponding cluster contents for further investigation. The user may choose to remove a comparison using the remove button on the right, in which case the table stats are recalculated and refreshed.

Columns 12 and 13, namely gas savings (kWh) and gas savings (%), are the energy savings that the adaptive mode achieves over legacy, in kWh and as a percentage respectively, for each pair. The table is sorted based on gas savings (kWh), but the user may choose to sort it based on any other column using the sorting controls. At the bottom of the table, with green text, the average daily savings in kWh and as a percentage are given. Note that for this report gas consumed for space heating only has been considered to calculate the savings.

Cluster 18

% gas saved — adaptive 4.41

legacy — 36.95

Search:

date	mean out temp	std out temp	mean target temp	mean target dist	std target dist	heat. demand (h)	boiler activ.	mean boiler temp	mean boiler max temp	bypass	avg heat. bal.	gas w/o DHW (kWh)	gas for DHW (kWh)	comfort
2021-03-11	9.57	2.46	19.99	0.58	1.11	18.97	8	45.73	57.38	adaptive**	4.41	35.48	1.66	59.70
2021-03-25	9.86	3.47	19.75	0.35	0.96	18.42	16	52.14	56.82	legacy	11.00	56.28	0.99	72.52

Cluster 12

% gas saved — adaptive 5.0

legacy — 9.63

Search:

date	mean out temp	std out temp	mean target temp	mean target dist	std target dist	heat. demand (h)	boiler activ.	mean boiler temp	mean boiler max temp	bypass	avg heat. bal.	gas w/o DHW (kWh)	gas for DHW (kWh)	comfort
2020-12-17	13.42	1.11	19.74	-0.32	0.91	18.01	35	39.74	55.09	legacy	11.0	33.06	2.18	83.6
2020-12-21	12.76	1.08	19.87	0.20	0.56	18.81	25	43.50	56.15	adaptive	5.0	30.23	0.31	86.1
2021-01-06	13.45	1.68	20.53	0.04	0.12	18.74	19	43.21	54.50	adaptive	5.0	29.51	1.38	100.0

Figure 6. Example clustering results.

The sections below the *Consumption Evaluation* report, provide more details regarding the contents of the generated clusters. The individual clusters are given, in tables, where the rows are referring to similar days that have been clustered together (Figure 6). A search box exists in each table that can filter rows with specific values (numerical or other) and the tables can be sorted by any column. In addition, a mini evaluation report for each cluster is given below the cluster title. It is made up by comparing the mean legacy performance of the cluster with all individual heating balance settings of the adaptive days. If no legacy day is found, the comparison is made with the highest heating balance setting. The reported values are the percentage of gas saved. Finally, all the days which don't match and can't be grouped with others are reported together in a separate section at the end.

4. Energy Disaggregation for Households

As more and more households are equipped with digital meters, the measured consumption profile could be used to profile the residents. Typical digital energy meters measure and report only active power, but the penetration of smart meters in households that measure both active and reactive power is increasing. Therefore, we have used this additional feature for more efficient detection of behind-the-meter devices.

Going one step further the metered data could be used to detect different activities or assets behind the meter. This can be further exploited to create more advanced digital twin models. Here we present the energy disaggregation method, also known as nonintrusive load monitoring (NILM).

4.1. Disaggregation algorithm

As shown in [5] , the disaggregation algorithm is based on detecting changes (events) in the household's real power (dP) and reactive power (dQ). By grouping and analyzing these dP-dQ based events, different NILM algorithms can be developed to detect different devices (Figure 7). The events are obtained from the analysis of the grid parameters voltage, real and reactive power and energy for all phases in the time domain with a high resolution in the range of seconds. The data are normalized to the nominal voltage [6] to filter out the dependence of the temporal grid characteristics on the device. For example, for most devices, a linear change in voltage would result in a quadratic change in power consumption. In addition, the very short power peaks (e.g., when the motor starts) are filtered out using nonlinear filtering methods to work only with the "stable" states. From the transformed data, the algorithm can detect simple short duration and long duration events. In addition, the algorithm attempts to distinguish concurrent events (e.g., a short duration event inside continuous event), which is critical for identifying or labeling all devices.

As shown in Figure 1, we implemented four different algorithms to identify and label the devices. Since there are some differences in the same device type, we tried to find the most unique events for each device type to avoid the need for long characterization of the devices or long training of the NILM algorithm.

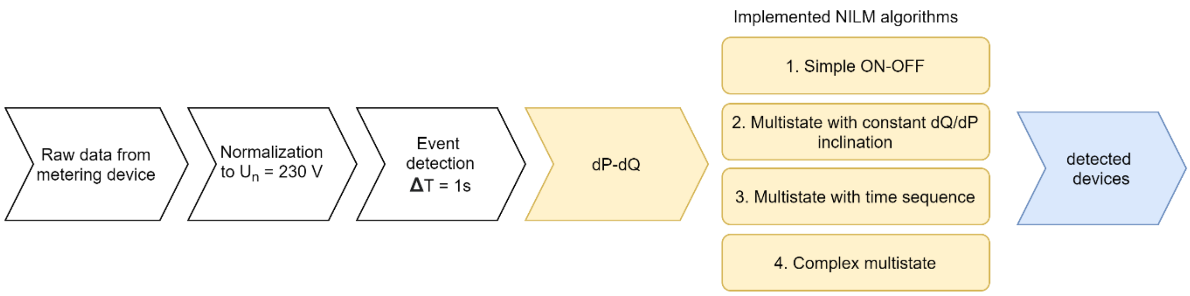


Figure 7. Flowchart of detecting events and using them as input for different NILM algorithms.

All four algorithms have a simple short description in the database of devices (Figure 8). Since different algorithms may consist of the same characteristic events, the order in the database is important and the first device in the base is analyzed first. The algorithms are:

1. **Simple ON-OFF.** This group includes simple devices with only one state (power characteristic). The most common devices are lights, irons, refrigerator/freezer, toaster, etc. The algorithm takes into account some small possible variations of the state, mostly caused

by the self-heating of the device, and therefore its characterization state may differ from its nominal power.

2. **Multistate with constant dQ/dP inclination.** These are typical devices that may have simple power control. The load characteristic inclinations (dP/dQ) are usually the same, regardless of consumption. Typical appliances are HVAC, cooktop, vacuum cleaner, ...
3. **Multistate with time sequence.** This group includes devices that have a typical pattern of states, but the length of the pattern and the time of occurrence of the states may vary. Typical devices from this group include dishwashers and washing machines.
4. **Complex multistate**. These are devices with many states that can occur at any time during operation. Characterization and detection of such devices may take more time compared to other methods. The most common representatives of this group are modern complex devices with implemented active power control such as TV, computers, ...

To improve detection, windowing in the time domain is used for device detection. The time window is five hours long and is determined by the longest unique multistate device (e.g., a washing machine). Windowing is very important to distinguish multistate devices from ordinary ON/OFF devices that may have a similar state to one of the multistate devices. The output of the algorithm is hourly data with total energy consumption and the labeled events. Since we use windowing, the data lags the current time by the size of the window.

```
description:fridge
type:ON/OFF
dP_ON : 70.0
dQ_ON : 20.0
dP_OFF :-70.0
dQ_OFF :-20.0
dPpercent : 45.0
dQpercent : 45.0
dtime : 3

description:HVAC
type:multistate_k
dP_ON_max : 2000.0
dQ_ON_max : 450.0
dP_ON_min :36.0
dQ_ON_min :9.0
dkPercent : 50.0
dtime : 50
```

Figure 8. A simple description of loads used in disaggregation algorithms.

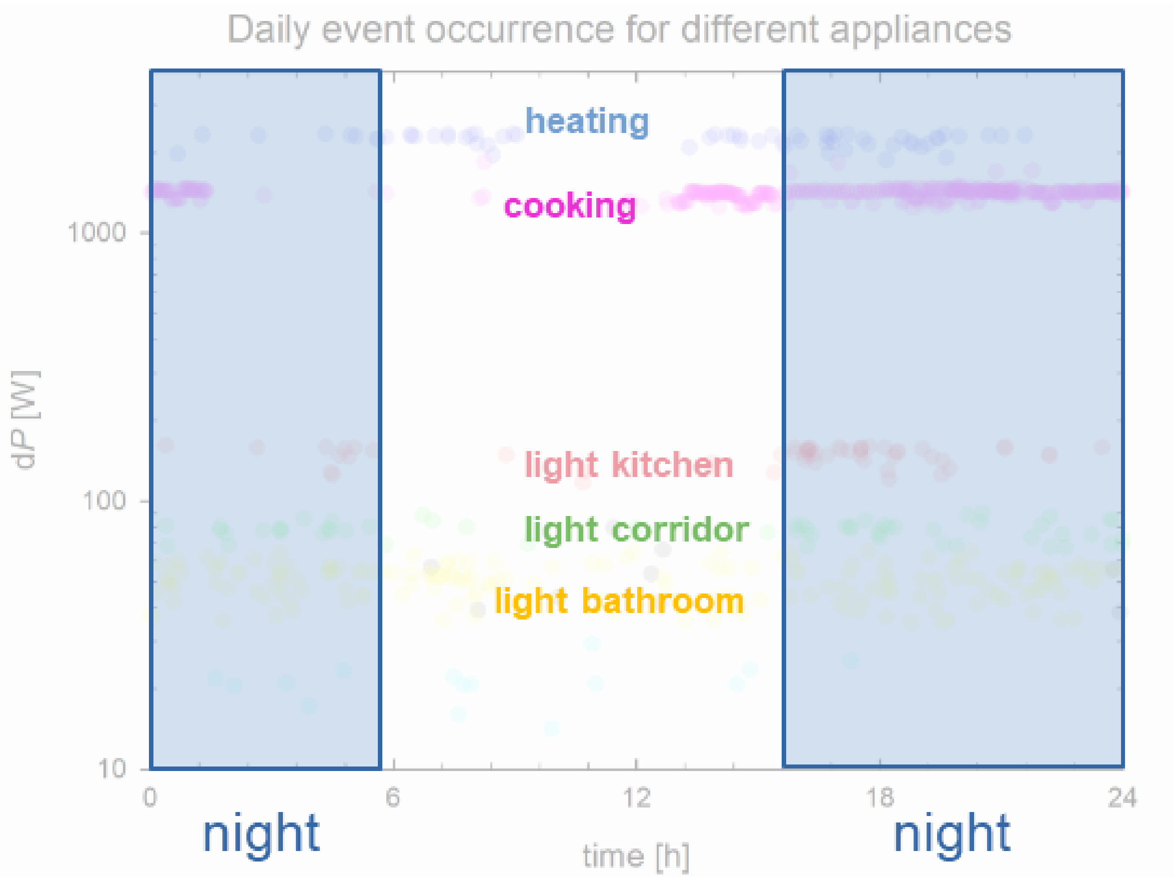


Figure 9. Typically measured power for one day (black solid curve) with detected and labelled events (coloured symbols).

In Figure 9 we see an example of the application of the developed algorithm to a household. We show the detection of devices on one of the phases of the household's three-phase system. There are some unlabeled events that are discarded. The events are colored according to the grouping. From the aggregated daily events of a week converted to a day, we can see the pattern of activities, a heat map of activities during a day. Once the results of the disaggregation algorithms are available as labeled data, various DTs can be created. For example, cooking-related activities could be easily filtered out and monitored to detect anomalies in cooking activities/behaviors.

5. Conclusions

This deliverable presented the first results on clustering and segmentation algorithms that are developed or are under development. These algorithms will improve the digital twins and forecasting algorithms developed within other WP4 tasks, and thus eventually also the flexibility services developed within WP5.

The deliverable presented 3 concrete approaches that are currently being investigated and developed:

- A clustering algorithm for residential gas boilers to assess the impact of an adaptive control strategy on gas consumption reduction and improved comfort for the end users.
- A proposed clustering methodology for electrical and thermal consumption profiles for a community of residential users, with as goal the refinement of the digital twins that are developed in T4.5 and demand response control strategies, developed in WP5.
- An ensemble of non-intrusive load monitoring algorithms for residential consumption profiles to identify the individual appliance consumption profiles. This provides insight in typical user lifestyle patterns and the potential flexibility from present controllable devices.

In a second version of this deliverable, planned for M30, an update will be provided on the presented algorithms, applied to (larger amounts of) BRIGHT pilot data and data from TNO's HESI lab, and incorporating the insights obtained from WP2 and WP3. We will also evaluate how these algorithms improve BRIGHT's digital twins and flexibility services.

References

[1] M. A. Zu, "Electric Water Heaters Using Dynamic Programming and K-Means Clustering," *IEEE Trans. Sustain. Energy*, vol. 11, no. 1, pp. 524–533, 2020.

[2] L. Diao, Y. Sun, Z. Chen, and J. Chen, "Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation," *Energy Build.*, vol. 147, pp. 47–66, 2017, doi: 10.1016/j.enbuild.2017.04.072.

[3] P. Gianniou, X. Liu, A. Heller, P. S. Nielsen, and C. Rode, "Clustering-based analysis for residential district heating data," *Energy Convers. Manag.*, vol. 165, no. December 2017, pp. 840–850, 2018, doi: 10.1016/j.enconman.2018.03.015.

[4] A. Javed, B. S. Lee, and D. M. Rizzo, "A benchmark study on time series clustering," *Mach. Learn. with Appl.*, vol. 1, no. May, p. 100001, 2020, doi: 10.1016/j.mlwa.2020.100001.

[5] G. W. Hart, "Nonintrusive appliance load monitoring," *Proceedings of the IEEE*, vol. 80, no. 12, pp. 1870–1891, Dec. 1992, doi: 10.1109/5.192069.

[6] M. Weiss, A. Helfenstein, F. Mattern, and T. Staake, "Leveraging smart meter data to recognize home appliances," in *2012 IEEE International Conference on Pervasive Computing and Communications*, Mar. 2012, pp. 190–197. doi: 10.1109/PerCom.2012.6199866.