energies

MDPI

Article

A Novel Feature Set for Low-Voltage Consumers, Based on the Temporal Dependence of Consumption and Peak Demands

Robbert Claeys *, Hakim Azaioud, Rémy Cleenwerck, Jos Knockaert and Jan Desmet

EELab/Lemcko, Department of Electromechanical, Systems and Metal Engineering, Ghent University, 8500 Kortrijk, Belgium; Hakim.Azaioud@UGent.be (H.A.); Remy.Cleenwerck@UGent.be (R.C.); Jos.Knockaert@UGent.be (J.K.); JanJ.Desmet@UGent.be (J.D.)
* Correspondence: Robbert.Claeys@UGent.be

Abstract: This paper proposes a novel feature construction methodology aiming at both clustering yearly load profiles of low-voltage consumers, as well as investigating the stochastic nature of their peak demands. These load profiles describe the electricity consumption over a one-year period, allowing the study of seasonal dependence. The clustering of load curves has been extensively studied in literature, where clustering of daily or weekly load curves based on temporal features has received the most research attention. The proposed feature construction aims at generating a new set of variables that can be used in machine learning applications, stepping away from traditional, high dimensional, chronological feature sets. This paper presents a novel feature set based on two types of features: respectively the consumption time window on a daily and weekly basis, and the time of occurrence of peak demands. An analytic expression for the load duration curve is validated and leveraged in order to define the the region that has to be considered as peak demand region. The clustering results using the proposed set of features on a dataset of measured Flemish consumers at 15-min resolution are evaluated and interpreted, where special attention is given to the stochastic nature of the peak demands.

Keywords: load profiling; consumer categorization; clustering; load duration curve; peak demand; feature construction

check for updates

Citation: Claeys, R.; Azaioud, H.; Cleenwerck, R.; Knockaert, J.; Desmet, J. A Novel Feature Set for Low-Voltage Consumers, Based on the Temporal Dependence of Consumption and Peak Demands. *Energies* **2021**, *14*, 139. https://doi.org/10.3390/en14010139

Received: 26 November 2020
Accepted: 18 December 2020
Published: 29 December 2020

1. Introduction

In different regions and countries in the European Union, including Flanders, the regulator for the electricity market has proposed an update to the traditional tariff structure for consumers connected to the low-voltage distribution grid [1,2]. The goal of this update is to obtain a tariff structure that better reflects the real costs associated with operating the distribution grid, as well as to incentivize consumers to change their consumption behavior. The emergence of digital meter technology and its rollout in European countries allows policymakers to implement these changes. Simultaneously, due to the higher time resolution, consumers are given a tool to gain additional insight in their consumption and related electricity invoice.

Residential and low-voltage consumers form a particularly challenging group from the viewpoint of grid operators and parties responsable for local grid balancing. Individual household consumption profiles are very behavior-dependent and often described as being peak-intensive and stochastic [3]. They often exhibit short peak demands, while simultaneously being characterized by large periods during the day and at night with very low energy demand. Therefore, regulators are proposing tariff structures that include cost elements related to both energy and capacity.

Consequently, the research attention has shifted to the mitigation of the peak demands by applying e.g., peak shaving programs or demand respons initiatives, as well as analyses on the predictability and stochasticity of these peaks. The research field related to customer categorization and load profiling aims to support policymakers and stakeholders by providing insights into (i) the types of consumers, and (ii) the behavior and differences between their load profiles. The traditional example of this is the construction of synthetic or representative load profiles. For residential consumers, the Flemish regulator currently offers two synthetic load profiles on an annual basis, i.e., households with and without electric heating.

In standard modeling techniques, the following steps are taken [4]. First, consumption profiles obtained from metering data with similar behavior have to be grouped together via a clustering algorithm. Unsupervised learning techniques are often used to detect underlying structures in large datasets, with K-means and hierarchical clustering among the most prevalent methods [5]. These algorithms can be performed either on the chronological data itself, or on a feature set obtained by transforming this chronological data. Common examples for chronological data include clustering based on daily or weekly profiles [6,7]. The seasonal influence can be taken into account by either performing a clustering process for each individual season [8], or by determining the recurring daily load profiles on an annual basis [9]. Typical load profiles can now be found by using statistical measures on the grouped chronological data, such as the mean or the median value at each time step [10]. This illustrates the main drawback during the construction of synthetic load profiles for purposes related to peak demands, such as emerging capacity-based tariffs. The averaging process results in a loss of important time-sensitive information unique to the individual household, and less volatile profiles are obtained [11].

As mentioned, an alternative method to clustering via the chronological measurements is grouping consumers based on similar properties, also called features. This work follows the feature construction and evaluation approach. The advantages of using a limited set of features during a clustering process are multifold. First, artificial overfitting due to high dimensional data can be avoided [12]. Furthermore, computational time is saved and allows easier interpretation if the features are chosen to be application-dependent [13]. Features can be constructed by performing operations on the default chronological features, e.g., combining all daytime consumption in one single feature. However, more advanced features can also be constructed, ranging from features generated in the frequency domain [14,15], to features related to the shape of the distribution of the load, such as the load factor [16].

Features constructed in literature are often application-dependent, i.e., depending on the goal of the work. One goal of this work is to investigate the temporal connection between consumption and peak demand behavior, to gain insight in the stochasticity of residential peak demands. Therefore, the features in this work are linked to either the consumption or the occurrence of peak demands. Previous studies incorporating temporal properties of these peaks in the clustering process either take the timing and the amplitude of the daily peak demand into account [17,18], or use statistical measures of the distribution of the measurement data [19]. In this work, the load duration curve (LDC), also called the demand frequency distribution graph, of each individual consumer is used to define which of its measurements constitute a peak demand on annual basis. The LDC is obtained by ordering the measurements in descending order. At the macrogrid level, the LDC has traditionally been used by electric utility engineers for network planning purposes, to analyse the utilisation of power plants, as well as characterizing the cyclic behavior of electricity demand [20–22]. While the LDC has not traditionally been used to model individual consumers, it was successfully used by Poulin et al. [23] to investigate the value of peak shaving for industrial and commercial consumers. Encouraged by these findings, the analytical form of the load duration curve is used in this work to construct the peak-based features. Based on the shape of the LDC, a threshold unique to each consumer is proposed, and every demand higher than this threshold can be considered as a peak. By determining the time of occurrence of these peak demands, the temporal properties of the peaks are taken into account.

The main contribution of this work thus is the introduction of features related to the peak demands. The introduced methodology combines both frequency- and time-based information to determine these features. The remainder of this work is organized as follows. In Section 2.1, the dataset and the preprocessing steps are described. Section 2.2 then proposes and validates an analytic expression for the load duration curves of low-voltage consumers. This expression is subsequently used in Section 2.3 to define the region that can be considered as peak demands. The features related to the timing of the consumption and peak demands are constructed in Section 2.4. This feature set is used for two purposes, and depending on the methodology, a different feature transformation has to be performed. The clustering algorithm, as well as the methodology used to analyse the stochastic nature of the peak demands, are described respectively in Sections 2.5.1 and 2.5.2. Section 3.1 reports on the findings and performance of the clustering algorithm, while Section 3.2 considers the relations between the consumption and peak demands in certain time periods to shed light on the stochasticity of these peak demands. Finally, Section 4 concludes this paper.

2. Materials and Methods

2.1. Consumption Profiles

The used dataset used in this work comprises 1422 consumers on the low-voltage distribution grid in two small Flemish towns in a suburban area, measured at a 15-min resolution during one year, leading to 35,040 time points per consumer. The data were provided by Fluvius cvba, the Flemish distribution network operator. The metering infrastructure was installed during a proof-of-concept study on digital meters in Flanders during the period 2010–2014. As more than 3000 households spanning different generations and compositions participated in this study, the dataset can be considered sufficiently diverse for consumers on the low-voltage distribution grid.

Several preprocessing steps were undertaken to obtain the final dataset, leading to a reduction from over 3.000 load profiles to 1422 data entries. These preprocessing steps are as follows:

- A first preprocessing step involving possible missing data was performed by the distribution network operator before providing the dataset for this research;
- Only meters that had measurements for the full year 2013 were included, given the purpose of this work;
- Households equipped with a PV installation were excluded from the analysis, as it is known that the presence of a PV installation can induce behavioral changes to increase PV self-consumption [24]. Furthermore, the metering data for households with PV installations merely included information on the net consumption and injection, not the gross consumption which is necessary for the proposed methodology;
- Following the Eurostat classification [25], meters indicating an annual consumption lower than 1000 kWh or higher than 15,000 kWh were excluded, as these were assumed to not be representative for typical household behavior, or could include small and medium-sized enterprises (SMEs), meaning commercial meters, on the low-voltage distribution grid.

After the preprocessing, the 1422 individual timestamped profiles are subsequently categorised based on the thermal images obtained via heatmaps of their demand profile. The introduced categories will not be used as input for the clustering algorithm, merely used for a post-hoc validation and interpretation of the obtained clusters in Section 3.1. This heatmap is the visualisation of the matrix obtained by reshaping the 35,040 ×1 vector of the chronological data to a 96 × 365 matrix. The entries belonging to the days of the start and end of daylight saving time are removed before reshaping the matrix, resulting in a 96 × 363 matrix. These days contain 92 and 100 data points, and would therefore distort the heatmap. Based on the obtained heatmaps, five categories are introduced that are able to describe the typical low-voltage consumers in Flanders: four behavior-specific categories and one so-called regular residential consumer for all consumers that do not fit one of the four special categories, inspired by the common Synthetic Load Profiles (SLPs) for Flanders.

The five categories are as follows:

- **SME profile**: consistent load profile with a 9–18 h behavior on weekdays and absent on weekends, as shown on Figure 1a;
- **Electric heating**: consumption late in the evening and at night, superimposed upon a regular consumption profile. Two substructures are observed:
  - **Ripple control heating**: These profiles exhibit the same moment during weekdays when the heating is turned on, and a different behavior is observed for weekdays and weekends, as shown on Figure 1b;
  - **Continuous heating**: Unlike the ripple control heating, the moment of switching on the heating is stochastic and no difference in heating behavior between weekdays and weekends can be observed, as shown on Figure 1c;
- **Air conditioning**: profiles with a significant electric load during summer months, superimposed upon a regular consumption profile. This heatmap is not shown for brevity.
- **Regular residential consumer**: the remaining load profiles not belonging to one of the above categories. There are typically (but not necessarily) characterized by a morning and evening peak, with demands concentrated during the evening as shown on Figure 1;

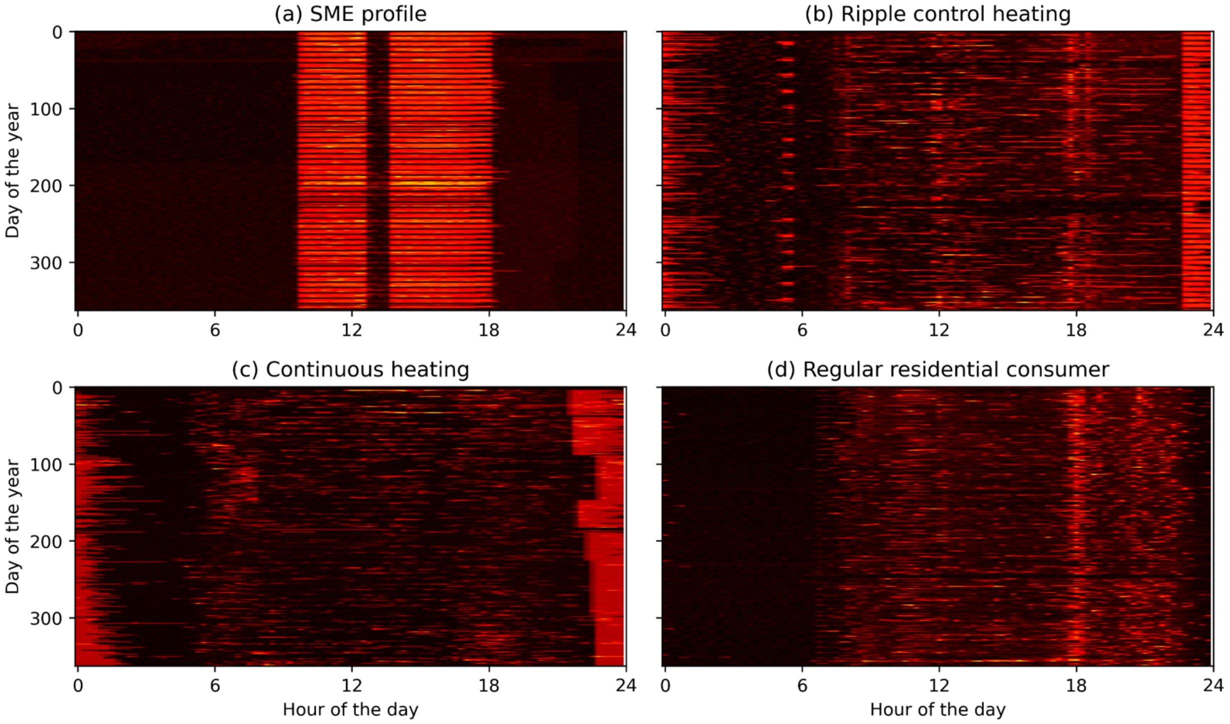

Figure 1. Examples of heatmaps for four different types of consumers on the low-voltage grid: (**a**) an SME profile, (**b**) a consumer with ripple control electric heating with a fixed start time of the heating, as well as differences between weekdays and weekends, (**c**) a consumer with continuous electric heating, and (**d**) a regular residential consumer.

As can be seen in Table 1, the majority of the considered consumers on the low-voltage distribution grid does not fall within a category with specific features such as the SME or the electric heating profiles, but can be considered a regular household. Table 1 gives an overview of the number of profiles in each category, split for different consumption ranges.

The density of the regular demand profiles is the highest in the range of 2–3 MWh per year, which is consistent with the most common household consumption in Flanders. Similarly, the other categories are mostly concentrated at higher average yearly consumption.

Table 1. Dataset composition by the defined profile categories, by consumption ranges.

Consumption Range	Regular	Ripple Control e-Heating	Continuous e-Heating	SME	Airco
1–2 MWh	143	3	1	0	0
2–3 MWh	260	19	6	2	0
3–4 MWh	254	25	4	2	0
4–5 MWh	223	17	15	0	0
5–6 MWh	126	14	13	5	0
6–7 MWh	86	10	14	1	3
>7 MWh	114	19	18	10	15
Total	1.206	107	71	20	18

2.2. Load Duration Curves

The load duration curve of an individual consumer is obtained by ordering its metering data in a descending order rather than the traditional chronological order. The analytic expression introduced by Poulin et al. [23] for commercial, institutional and industrial consumers is taken as the starting point for the analysis on the low-voltage consumers considered in this work. Let $P^i(t)$ denote the chronological demand data of a specific consumer i, its corresponding LDC $\mathcal{P}^i(\tau)$ can subsequently be written as:

$$\mathcal{P}^i(\tau) = 1 - a\tau - b\tau^c + \frac{d}{1 + e^{f(\tau - g)}} - \frac{d}{1 + e^{fg}} \tag{1}$$

The variables τ and $\mathcal{P}^i$ in the expression of the LDC respectively denote the normalized time and normalized demand, i.e., both scaled such that their range spans the interval $[0, 1]$. This allows for a scale-independent comparison between consumers, merely comparing the behavior of the demand curves. The six parameters included in Equation (1) show a clear connection to customer operations, and thus are relevant for consumer clustering purposes. The peak height and duration are correlated with b and c respectively, while parameters d, f and g are linked to respectively the height, slope and location of the step. Finally, a yields information about the general slope of the curve. These six parameters and their relation to the general shape of the LDC given by Equation (1) are given in Figure 2a.

While the six-parameter expression was previously validated to accurately model the LDC of individual and aggregated residential consumers [26], this work aims to both simplify this six-parameter expression, as well as to link the parameters in its simplified expression to properties of the consumer, such as the annual consumption. Therefore, Figure 2b displays the shape of the first proposed improvement for low-voltage consumers, a 5-parameter model. Intuitively one could indeed expect households to spend the majority of their year on a certain baseload, i.e., the aggregated standby demand of the appliances in the household. As such, this would correspond to a saturation effect towards this standby demand being present in the household LDC for $\lim_{\tau \to 1} \mathcal{P}^i(\tau)$, in contrast to the decreasing slope that is present in Equation (1).

Consequently, a five-parameter LDC model for low-voltage distribution grid consumers is proposed in Equation (2). The linear term included in Equation (1) is omitted.

$$\mathcal{P}^i(\tau) = 1 - b\tau^c + \frac{d}{1 + e^{f(\tau - g)}} - \frac{d}{1 + e^{fg}} \tag{2}$$

The LDC based on the measurement data is constructed for each individual metering point and the values of the parameters included in Equations (1) and (2) are subsequently determined via a curve fitting algorithm. This procedure is performed via the `lmfit` package in Python, where a Least-Squares minimization with a Trust Region Reflective method is used [27].

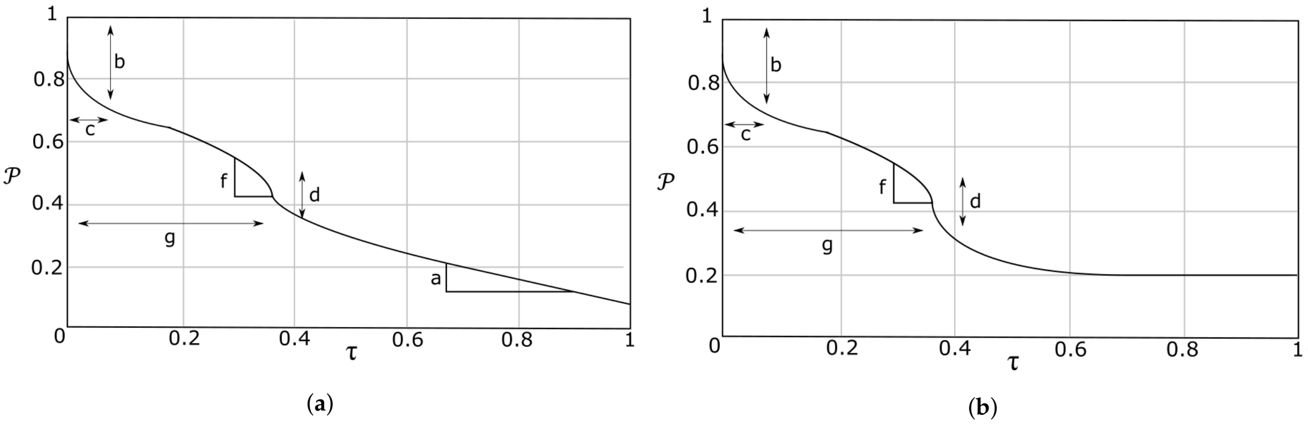

Figure 2. (**a**) Shape of the 6-parameter LDC for commercial, institutional and industrial consumers [23], and (**b**) shape of the proposed 5-parameter LDC for low-voltage consumers.

The parameters are constrained according to following boundaries:

$$\begin{cases} 0 \leq a, b, c \leq 1 \\ 0.02 \leq d \leq 1 \\ 25 \leq f \leq +\infty \\ 0.025 \leq g \leq 1 \end{cases}$$

The boundaries were chosen nearly identical to those used for the LDC fitting procedure in [23]. The only deviation from the boundaries in [23] is the lower bound of the g parameter. Therein, a lower bound of 0.1 for g was assumed. However, as mentioned in Section 1, residential consumers are more peak-intensive and their peaks are more stochastic. It is expected that this behavior is reflected in the shape of the LDC with a shorter duration of the peak and step. Therefore, the lower bound for g, the parameter linked to the location of the step, can be taken smaller than the aforementioned 0.1. A value of 0.025 was chosen for this lower bound.

On average, the correlation coefficient R^2 is 0.977 for both expressions. However, for 91% of the consumers in the dataset, both the Akaike and the Bayesian information criterion point toward Equation (2) as the most suitable expression to describe the LDC. This is further supported by the very small value of a for the 6-parameter expression: on average, a has a value of 0.0026, corresponding to a near negligible slope in Equation (1).

However, describing consumers via their load duration curve has several disadvantages. One inherent disadvantage of the load duration curve is the loss of all temporal information, which forms the subject of Section 2.4. A second disadvantage is related to the use of the expression for the demand-normalized LDC. While this allows for a more straightforward and scale-independent comparison of the parameters describing the behavior of individual consumers, all information related to the original peak demand is lost, and no information on the traditional properties such as the annual consumption is retained.

Therefore, the second step of this analysis entails incorporating possible correlation between the values of the parameters in Equation (2) and the annual consumption. The relation between the individual parameters of the LDC and the yearly consumption of the consumer is given in Figure 3. Despite a large spread being present in the scatterplot, the parameter c describing the power law in Equation (2) is noticeably correlated with the yearly consumption. Consequently, Equation (3) fixes the parameter c at the value $c_0 + c_1 Y$. The values of c_0 and c_1 are determined by an ordinary least-squares fitting procedure on the relation between Y, the yearly consumption in kWh, and c, as shown in Figure 3.

$$\mathcal{P}^i(\tau) = 1 - b\tau^{c_0 + c_1 Y} + \frac{d}{1 + e^{f(\tau - g)}} - \frac{d}{1 + e^{fg}} \tag{3}$$

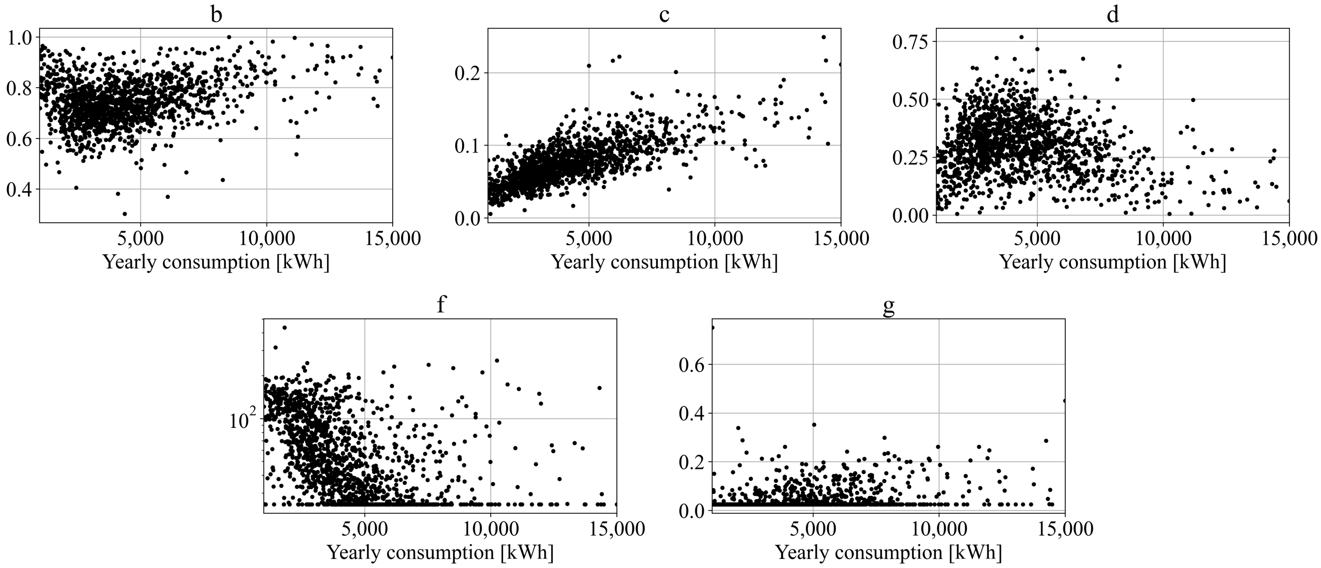

Figure 3. Scatterplots of the parameters included in Equation (2), plotted versus the yearly consumption.

With an obtained value of 0.0345 and 0.456×10^{-6} for c_0 and c_1 respectively, an analogous fitting procedure of the LDC to Equation (3) is performed. The result and a comparison with the previous two models is listed in Table 2. While a decrease in median R^2 value can be observed, this value is still acceptable. However, the observed mean value is significantly lower and exhibits an increasing difference with the median value, highlighting that the 4-parameter model of Equation (3) leads to a worse fit for a non-negligible amount of consumers.

Table 2. Comparison of the fitting result of the three considered LDC models.

Model	Median R^2 Value	Mean R^2 Value
6-parameter model, Equation (1)	0.987	0.977
5-parameter model, Equation (2)	0.987	0.977
4-parameter model, Equation (3)	0.968	0.937

As expected, given the high spread in the linear relation between c and the yearly consumption, the reduction in accuracy of modeling the LDC is a trade-off that has to be made in order to incorporate the dependency on the consumer's yearly consumption. Given the importance of the fitted parameters of the LDC in the remainder of this work, the further analyses are performed on Equation (2), the model that exhibited superior performance in the fitting procedure.

2.3. *Definition of Peak Demands*

The validated analytic expression of the load duration curve can now be used to introduce a binary classification for peak demands, i.e., all values $\mathcal{P}^i(\tau)$ for τ smaller than a certain threshold τ^* can be considered peaks for the individual consumer while all other values cannot. The challenge now lies in determining τ^*, the value of this threshold. The only condition a proposed expression or value for τ^* has to fulfill for the purposes intended in this work is that it has to be sufficiently small in order to yield usable results. Although the term "usable" implies a certain level of arbitrariness, it should be clear that a threshold value that labels 50% of all demands on yearly basis as peaks is not practical for e.g., peak shaving algorithms. Therefore, given the continuous nature of the load duration curve, it is inevitable that any proposed threshold value will have its own advantages and disadvantages.

This work proposes using the point of maximum curvature as this threshold, for the function for τ sufficiently small. Intuitively, the curvature of a function is the amount by which this function deviates from a straight line in a certain point. Therefore, the maximum of this curvature function denotes the point where the curve has the sharpest bend.

For τ sufficiently small, the LDC as defined in Equation (2) can be approximated by Equation (4), which is dominated by the power law responsible for the peak demand features and the steep decay of the LDC:

$$\mathcal{P}^i(\tau) \approx 1 - b\tau^c. \tag{4}$$

Using the point of maximum curvature of Equation (4) as the threshold value to define the area of peak demands has two major advantages. First, this threshold is different for each individual as it depends on the shape of the individual load duration curve, allowing for a differentiation among low-voltage consumers. Second, the point of maximum curvature for an analytic function can be unambiguously described analytically. The curvature function $\kappa(\tau)$ of Equation (4) is given by:

$$\kappa(\tau) = \frac{\left|(\mathcal{P}^i)''(\tau)\right|}{\left[1 + \left[(\mathcal{P}^i)'(\tau)\right]^2\right]^{\frac{3}{2}}} \tag{5}$$

Maximizing $\kappa(\tau)$ with respect to τ yields following value for the point of maximum curvature:

$$\tau^* = \left(\frac{c-2}{b^2c^2(2c-1)}\right)^{\frac{1}{2(c-1)}} \tag{6}$$

The histogram of the calculated values of τ^* and the corresponding value $\mathcal{P}^i(\tau^*)$ for the considered dataset is given in Figure 4. A beta probability density function is successfully fitted and shown to be able to describe the density functions, as shown overlaid in Figure 4. The distribution of τ^* has a 10–90 percentile range of [0.017, 0.041], with a mean value of 0.028. Translating this mean value of the normalised time τ to a yearly basis means that, on average across the distribution, 2.8% of the values on a yearly basis can be labeled as peaks, corresponding with 981 values of the 35.040 data points. Furthermore, the distribution of $\mathcal{P}^i(\tau^*)$ shows the large potential of peak shaving initiatives: the mean value of $\mathcal{P}^i(\tau^*)$ is 0.35, i.e., 35% of the original maximum demand.

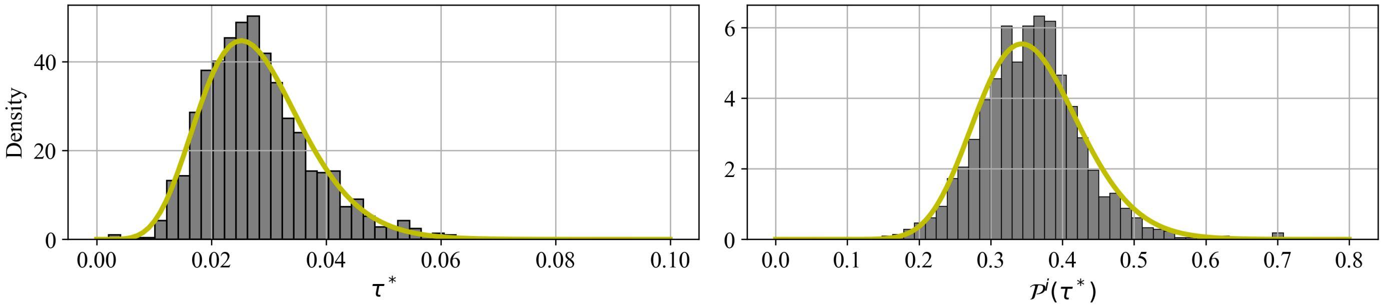

Figure 4. Density histogram of the calculated values of τ^* and $\mathcal{P}^i(\tau^*)$, as defined by Equation (6).

2.4. Feature Construction

The aim of this work is to construct a feature set that can be used for dual purposes. First, to group similar consumers together in so-called clusters based on properties that are relevant for (i) distribution network operators with respect to their operating and tariffing purposes, as well as for (ii) individual consumers, i.e., for the local applicability of distributed energy resources and how evolving tariff structures impact them. Second, to investigate the stochastic nature of peak demands, given the current trend of countries to introduce capacity-based tariff schemes for low-voltage consumers combined with the uptake of small-scale storage systems that can be used for peak shaving purposes in residential areas. Given the intended purposes and the large temporal dependence of renewable yields, we choose to explicitly incorporate the temporal influence on the demand profile in the construction of the feature set.

Two different temporal levels relevant for low-voltage distribution grids are considered for these time-dependent features: the daily and the weekly level. At the daily level, the time periods are defined based on the time of day. Similarly, at the weekly level the distinction between weekday and weekend is maintained. In order to distinguish between intervals $\mathcal{I}$ defined on either the daily level and the weekly level, two notations are introduced: $\mathcal{I}^d$ and $\mathcal{I}^w$. The superscript d stands for daily, w for weekly.

While the definition of the intervals $\mathcal{I}^w$ is unambiguous, i.e., weekdays versus weekends, distinct intervals at the daily level $\mathcal{I}^d$ for residential consumers are not universally agreed upon. In [17], Haben et al. identified four key time periods for residential consumers: overnight, breakfast, daytime and evening period. Inspired by their findings, the distinction as listed in Table 3 is introduced for the daily level. In this work, the daytime period is further subdivided in a morning and afternoon range. Furthermore, while the daytime period in Ref. [17] ended at 15:30, it is extended to 18:00 for this work.

Table 3. Definition of the considered time periods $\mathcal{I}^d$ at the daily level, based on the hour of the day, h.

$\mathcal{I}^d$	**Definition**
Early morning	$h \in$ [06:00–08:30]
Morning	$h \in$ [08:30–12:00]
Afternoon	$h \in$ [12:00–18:00]
Evening	$h \in$ [18:00–22:30]
Night	$h \in$ [22:30–06:00]

Other temporal levels can easily be incorporated in the feature set, e.g., the seasonal influence by including four time periods at the annual level corresponding with the seasons. However, this seasonal variation is omitted in this work, as these features were found to not significantly impact the clustering result in Section 3.1 and rather obfuscated the results, limiting the ease of interpretation.

Based on these time periods at two different temporal levels, a two-pronged approach is introduced in the subsequent subsections. The first class of features, which will be discussed in Section 2.4.1, considers the relation between the temporal property and the consumption: which fraction of the demand occurs during a certain predefined time interval? In contrast, the second class of features, discussed in Section 2.4.2, considers the temporal properties of the peak demands: when do these peak demands occur? In order to unambiguously define which values constitute a peak, the analysis performed on the analytic form of the load duration curve in Section 2.3 is used.

The individual features are suitable to characterize consumers, e.g., for assessing household compatibility with renewable energy sources (households with high daytime consumption are more favorable for rooftop-integrated PV installations without a battery), or for the timing of the individual peak demands, which is beneficial information for distribution network operators. However, it is the knowledge on the fraction of the demand combined with the simultaneous occurence or absence of peak demands in that time period that can clarify the stochastic nature of these peak demands. Consumers that consistently exhibit a disproportionate amount of peak demands in a certain time period can be targeted for peak shaving initiatives, either via demand response programs or by utilising an energy storage system.

2.4.1. Temporal Dependence of Consumption

Let $\mathcal{B}^i_{\mathcal{I}^x}$ be the subset of all measured values $P^i(t)$ of consumer i that occur in one of the previously defined time periods $\mathcal{I}^x$, with the superscript x denoting the considered temporal level. This yields following definition of this subset:

$$\mathcal{B}^i_{\mathcal{I}^x} = \left\{ P^i(t) \mid t \in \mathcal{I}^x \right\}, \quad x \in \{d, w\} \tag{7}$$

The fraction $f^{i,c}_{\mathcal{I}^x}$ of the demand of consumer i in time period $\mathcal{I}^{x\,i}$ is given by:

$$f^{i,c}_{\mathcal{I}^x} = \frac{\sum_{y \in \mathcal{B}^i_{\mathcal{I}^x}} y}{\sum_t P^i(t)} \tag{8}$$

This definition yields a total of seven features: five features for the daily level, two for the weekly level. However, as the subsets $\mathcal{B}^i_{\mathcal{I}^x}$ for a given temporal level x are disjoint by construction, the sum of $f^{i,c}_{\mathcal{I}^x}$ over all $\mathcal{I}^x$ for a fixed x is equal to 1. Therefore, this reduces down to five linearly independent features: four for the daily level, one for the weekly level.

2.4.2. Temporal Dependence of Peak Demands

The features related to the peak demands are treated in a different way than those linked to the consumption. While the amplitude of the demand $P^i(t)$ at a certain point in time is important to determine the fraction of consumption that happens in a time interval, only the presence of peak demands is of importance for the second set of features, not the size of the peaks. Let $\mathcal{D}^i$ be the subset of all measured demand values $P^i(t)$ of consumer i that can be considered as a peak demand, as defined in Section 2.3. As the LDC is normalized with respect to the annual peak demand $P^i_{\max}$, the value $\mathcal{P}^i(\tau^*)$ has to be rescaled:

$$\mathcal{D}^i = \left\{ P^i(t) | \, P^i(t) \geq \mathcal{P}^i(\tau^*) \cdot P^i_{\max} \right\} \tag{9}$$

Analogous to the previous section, let $\mathcal{D}^i_{\mathcal{I}^x}$ now be the subset of $\mathcal{D}^i$ that occurs in time period $\mathcal{I}^x$:

$$\mathcal{D}^i_{\mathcal{I}^x} = \left\{ P^i(t) \mid P^i(t) \geq \mathcal{P}^i(\tau^*) \cdot P^i_{\max} \wedge t \in \mathcal{I}^x \right\}, \quad x \in \{d, w\} \tag{10}$$

The number of peak demands per time interval can be found via the cardinality of the set $\mathcal{D}^i_{\mathcal{I}^x}$, i.e., $\left|\mathcal{D}^i_{\mathcal{I}^x}\right|$. The fraction of peak demands for consumer i in a certain time period, $f^{i,p}_{\mathcal{I}^x}$, can therefore be found via Equation (11). Analogous to the features related to the temporal aspect of the consumption behavior, this leads to another five linearly independent features. Consequently, this brings the number of considered linearly independent features for the clustering algorithm up to ten parameters.

$$f^{i,p}_{\mathcal{I}^x} = \frac{\left|\mathcal{D}^i_{\mathcal{I}^x}\right|}{\left|\mathcal{D}^i\right|} \tag{11}$$

These 14 features can now describe the temporal behavior and distribution of the consumption and peak demands, as illustrated for one randomly chosen regular household, household 802, in Figure 5. Both the fraction of the consumption and the fraction of peaks are shown for each time period at the daily and the weekly level. Major differences between the distribution describing the consumption and peaks can be observed. At the weekly level, 65% of the household's peaks are observed in the weekend, while only 35% of the consumption occurs during weekends. Similarly, more than 25% of consumption for this consumer happens at night, as defined by Table 3, while 10% of the peak demands lie in this time period. It is this difference between distributions of consumption and peak behavior at the same temporal level that forms the subject of the following sections, as the presence or absence of differences can clarify whether or not peak demands tend to be more stochastic or more deterministic.

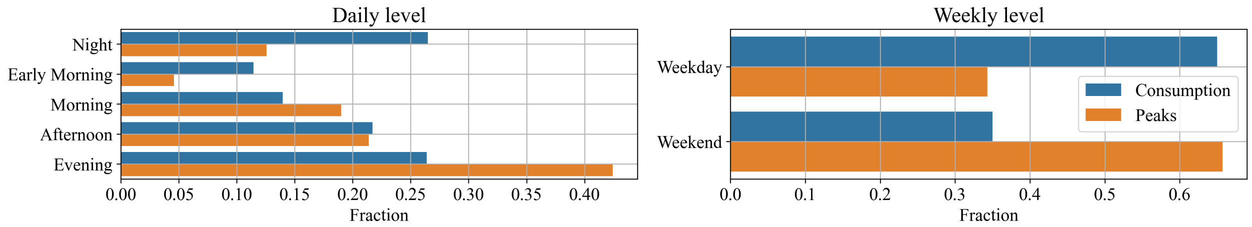

Figure 5. Example of the 14 features describing the temporal behavior of the consumption and peak demands at the daily and weekly level of household 802 in the dataset.

2.5. Transformation & Analyses on the Feature Set

One additional step is performed on the proposed set of features before proceeding to either the clustering algorithm or the distributional analysis: the feature transformation. Depending on the proposed methodology for each analysis, a different feature transformation is more appropriate. Therefore, this section discusses both the proposed methodology for each performed analysis, as well as the corresponding most suitable feature transformation.

2.5.1. Clustering Algorithm

No additional information or metadata is included in the dataset of load profiles. As such, the true underlying structure or the optimal amount of clusters to segmentate the dataset into is unknown. Therefore, unsupervised machine learning is used to cluster those profiles that exhibit similar behavior. The majority of the rich body of literature available on the unsupervised clustering of load profiles, whether chronologically ordered profiles or based on a constructed feature set, is based on one of two techniques: either a variant of the K-means clustering algorithm or via agglomerative clustering. In the structured literature review on the classification of consumption profiles performed by Tureczek and Nielsen, 65% of the considered papers included a K-means-based method, while another 29% performed analyses based on agglomerative clustering [4].

There are several differences between K-means and agglomerative clustering, both from a conceptual viewpoint, as well as the computational aspect. Agglomerative clustering offers a visualisation in a so-called dendrogram of the clustering results, intuitively showing how substructures in the dataset emerge when dividing or merging clusters. Furthermore, when a feature set is used as input for the agglomerative clustering, further analysis on the merging of clusters offers the possibility of tracking which features are the driving force that distinguish clusters. However, agglomerative clustering is a so-called greedy algorithm: at each step, the two closest clusters as defined by a linkage method are merged together. Therefore, agglomerative clustering techniques are prone to yield a sub-optimal solution instead of a global optimum. In contrast, given an input k, the number of desired clusters, a K-means algorithm partitions the dataset into k clusters. However, K-means tends to get stuck in a local minimum instead of the global minimum. The main challenge for a K-means approach lies in finding the optimal amount of clusters. From a computational point of view, K-means is preferable for larger datasets as the time complexity for K-means algorithms typically is linear in the input data size, $\mathcal{O}(n)$, while the time complexity for agglomerative clustering is quadratic, $\mathcal{O}(n^2)$.

In this work, an agglomerative clustering algorithm with Ward's linkage method is used, as implemented in Python's `scipy` package [28,29]. The main contribution of this work is introducing and validating a novel feature set. Therefore, the visualisation and emergence of substructures in the dataset in the clustering process is of major importance, justifying the choice for an agglomerative clustering algorithm. The proposed linkage method minimizes the total within-cluster variance for each merging step, which tends to yield approximately equally sized clusters. Following the arguments presented by Kang and Lee in [30], it is a necessary condition for clusters to have a roughly equal size, in order to be useful in real life applications according to expert opinions. Therefore, Ward's linkage method can be deemed appropriate, as the tendency of clustering algorithms to propose singular clusters that contain outliers is avoided.

For the proposed feature set, Ward's linkage method for agglomerative clustering relies on the Euclidean distance between the 10 linearly independent features in the 10-dimensional feature space. Therefore, obtained results will depend on the scale of the input features. However, when looking at both Table 3 and Figure 5 it is clear that the proposed features are not yet at the same scale.

By construction, the proposed time periods are not of the same scale, e.g., the weekend period is not the same length as the weekday period, nor is the early morning of similar length as the night interval. Therefore, even a uniform distribution would not lead to similarly scaled features, leading to a distortion of importance of several features.

Therefore, an initial transformation is performed that rescales the features based on the length of their time period such that in the case of a uniform distribution, the value of all features $f^i_{\mathcal{I}^x}$ would be equal to 1. Any deviation of a uniform distribution will then lead to a deviation of this unity value for each parameter, while avoiding an artificial inflation of the importance of an individual feature or one set of features. However, of the ten proposed linearly independent features, eight are defined on daily basis, while only two are defined on weekly basis. While this initially proposed transformation aims to give each individual feature the same weight, the two sets of features defined on different temporal levels are not a priori equally represented in the feature set. Consequently, instead of transforming the features on weekly basis to be equal to 1 in the case of a uniform distribution, they are assigned an additional weighting factor equal to 2 to partially offset the numerical advantage of daily features.

In summary, the two sets of features proposed in Section 2.4 are transformed in a two-step transformation before being used as input for the hierarchical clustering, using a Ward's linkage method. First, the features are rescaled based on the length of the time interval in which they are defined, which leads to individual features of the same scale. In the second step, an additional weighting factor is assigned based on the amount of features for each temporal level. A weighting factor of 2 is proposed for the weekly-level features, which partially offsets the numerical advantage daily-level features have in the proposed feature set. Further increasing this weighting factor would put a higher emphasis on the difference between weekdays and weekends in the clustering algorithm.

2.5.2. Distribution Analysis

The distribution of features $f^i_{\mathcal{I}^x}$ at the daily or weekly level x can yield interesting information. As mentioned before, households with high daytime consumption are ideal candidates for PV installations, whereas households that exhibit a large amount of peak demands in a certain time interval, could be targeted via demand response initiatives. However, it is the difference between the distributions describing the consumption and peak behavior at the daily or weekly level that yields information about the disproportionate presence of peak demands at a certain time interval, and thus about how stochastic the presence of peak demands are for an individual household. Therefore, two measures are proposed to investigate these distributions.

At the level of the individual distributions, we propose using the concept of entropy at the daily or weekly level to characterize the variability of household behavior. Similar to the goal of this work, Ref. [9] introduced entropy to study the variability of households, not with respect to features based on consumption of peak behavior, but based on the variability of consumption behavior described by the frequency of different representative daily load shapes during the year. Shannon entropy as introduced in information theory is defined in Equation (12), with x_i being a possible outcome and $p(x_i)$ the probability of this outcome [31].

$$H(x) = -\sum_{i=1}^{k} p(x_i) \ln p(x_i) \tag{12}$$

This definition of entropy has several interesting properties for this research. First, in the case of a uniform distribution, the entropy reaches its maximum and thus yields maximum uncertainty. Second, any deviation from a uniform distribution results in a decreasing entropy, and thus less uncertainty. If there is no uncertainty, then the entropy becomes 0. In order for these properties to hold in the analysis of the introduced features, it is important that these features are consistent with the assumptions in the Shannon entropy. First, in the case of a uniform distribution, the entropy becomes maximal. A uniform distribution for the consumption behavior would entail having the same consumption at each time period.

However, as mentioned in Section 2.5.1, the periods as defined in Section 2.4 are not of equal length, which leads to inequal features in the case of a uniform consumption distribution. Therefore, the features are rescaled based on the length of the interval for which they are defined such that a uniform consumption distribution leads to identical consumption-related features. Furthermore, Equation (12) is defined for probabilities $p(x)$. As such, the features defined on the five periods at the daily level are rescaled to 0.2, while those defined at the two periods at the weekly level are rescaled to 0.5.

The second part of the distributional analysis entails a comparison between the distributions of the consumption behavior and the occurrence of peak demands at the daily and weekly level. As such, a measure for the distance between these two distributions has to be introduced. In this work, the Wasserstein-1 distance is used to characterize the distance metric between two probability distributions [32]. Here, the formal definition of the Wasserstein-1 distance as integrated in Python's `scipy` package is used [29]:

$$l_1(p,q) = \inf_{\pi \in \Gamma(p,q)} \int_{\mathbb{R}\times\mathbb{R}} |x-y| d\pi(x,y) \tag{13}$$

Here, p and q are two distribution functions, and $\Gamma(p,q)$ is the set of probability functions on $\mathbb{R} \times \mathbb{R}$ whose marginals are p and q on the first and second marginals respectively. This Wasserstein distance is also commonly called the earth-mover's distance, as it originated in the field of optimal transport issues. Intuitively, it can be seen as the minimum amount of "work" that has to be done to transform one distribution into the other, if each distribution could be considered as a pile of earth. The "work" takes into account both the distance it has to move, as well as the amount of earth it has to move. As such, distributions P and Q that are different over "long" (horizontal) regions will be far away from each other in the Wasserstein distance sense [33].

It is this property of the Wasserstein distance that is appropriate for this work. As the time periods at the daily level were introduced in an ad hoc way in Table 3, a distance metric that takes the horizontal difference into account instead of performing a pointwise comparison partly compensates the arbitrary nature of the definition. For a given consumption behavior, this allows us to identify distributions of the peak demands that are closer in time. This property is illustrated in Figure 6, where an artifical distribution of the consumption at the daily level is compared with two peak demand distributions. As the first probability distribution of the peak demands has a maximum chronologically closer to that of the normalised consumption, the Wasserstein distance is lower. The chronological ordering of the time periods as given in Figure 6 is chosen for the remainder of this analysis.

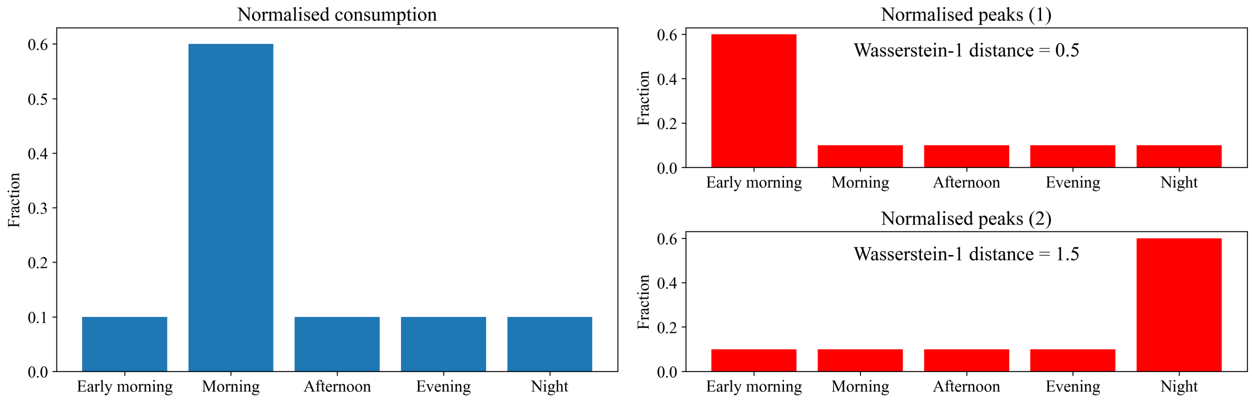

Figure 6. Illustration of how the Wasserstein distance is able to capture chronological differences between the probability distributions of consumption and peak demands.

3. Results

First, the results for the agglomerative clustering algorithm are illustrated based on the calculated dendrogram. The results for a low number of features are benchmarked to the available synthetic load profiles in Flanders, subsequently highlight how differences in feature behavior lead to the emergence of distinct and compact clusters, and argue how this knowledge can be leveraged from the viewpoint of demand response programs or peak shaving initiatives. Second, the distributions of the features at the same time levels are analysed. On the one hand, the Shannon entropy is used to characterize the variability of each type of feature. On the other hand, the Wasserstein-1 distance is used for an in-depth analysis of the stochastic nature of the peak demands, by comparing the distributions describing the household consumption and peak demand behavior.

3.1. Clustering Result

The dendrogram visualizing the hierarchical clustering process using Ward's linkage method on the proposed feature set is shown in Figure 7. Two horizontal cuts are included in the figure. The black line at $y = 30$ denotes the height where three clusters are obtained. This can serve as an initial benchmark, as there are three synthetic load profiles available for low-voltage consumers in Flanders: residential with and without electric heating, and non-residential. The red line was chosen such that 10 disjoint clusters emerge, leading to the color threshold of the highlighted clusters in the dendrogram. This threshold of 10 clusters was chosen based on two independent studies stating that for practical considerations, the total number of clusters should not exceed 10 [13,30]. This argument is based on the opinions of industrial experts, as these clusters are often used for tariffing or marketing purposes.

First, it is necessary to benchmark the clustering result to the available residential SLPs in Flanders. As the color threshold and further discussion in this section is based on 10 clusters, the analysis is performed based on the highlighted 10 clusters. By tracking the merging clusters into the three branches of the dendrogram at the cut $y = 30$, a benchmark can be performed. Figures 8 and 9, displaying respectively the distributions of the 14 untransformed features for the individual clusters, and the distributions of the yearly consumption of the consumers assigned to each clusters, allow for an interpretation of the obtained clusters based on consumer properties.

The first branch separates into clusters 1–3, the second into clusters 4–5, while the final branch leads to clusters 6–10. The clusters originating from these three branches are partitioned by dashed lines in Figure 8 for an easier comparison.

The following discussion on the benchmarking of the results is based on the observed feature distributions in Figure 8. The first branch groups consumers with a high fraction of consumption and peaks in the evening, which is typical for regular households. The second branch, containing clusters 4–5, groups consumers with a high fraction of the consumption and peaks at night. This is encouraging, as this could indicate the presence of electric heating, one of two major categories of residential consumers.

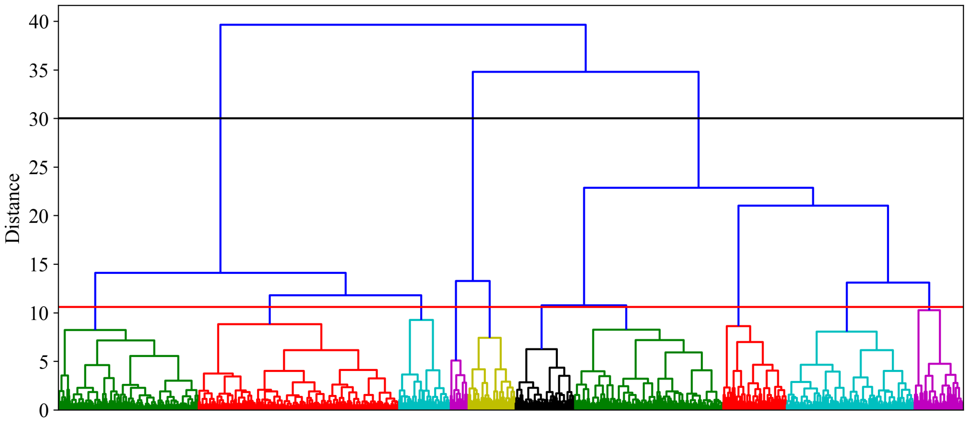

Figure 7. Dendrogram obtained via hierarchical clustering, with a color threshold highlighting 10 clusters. Individual profiles are given on the x-axis, while the y-axis denotes the distance.

The interpretation of the third branch is less straightforward, as the properties of the clusters composing this branch are more diffuse: (i) clusters 6–7 group consumers with a disproportionate amount of peaks during the weekend, (ii) cluster 8 collects the consumers with a significant amount of peaks during the early morning, whereas (iii) clusters 9–10 exhibit a large number of peaks during the morning and afternoon.

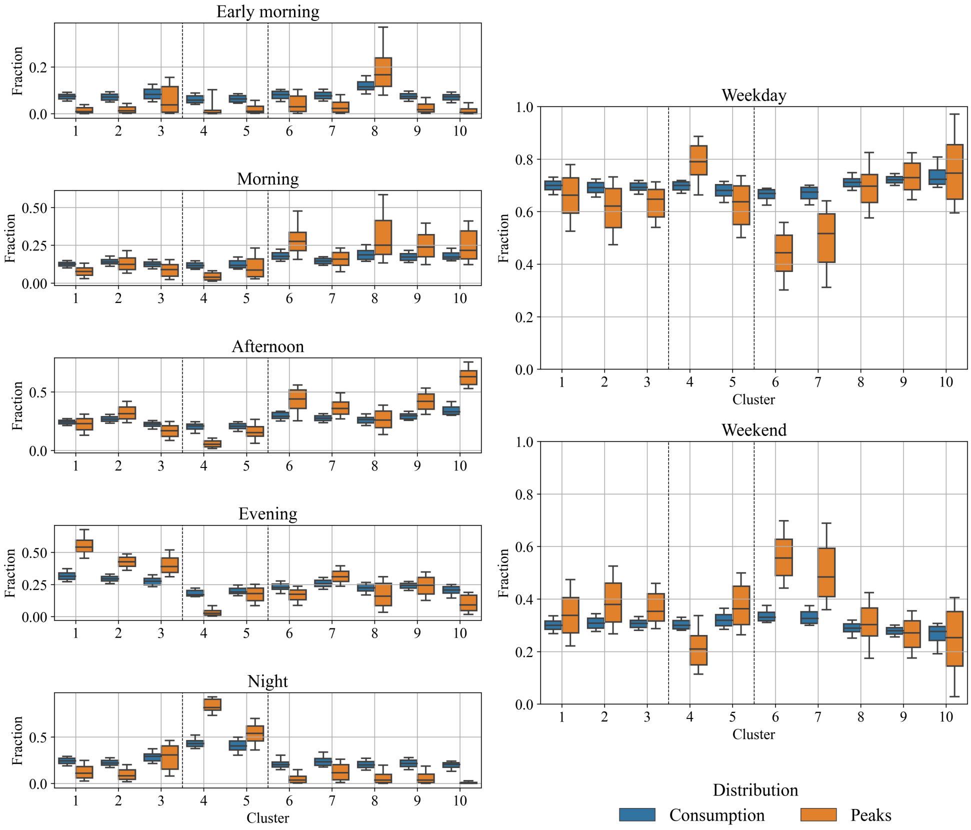

Figure 8. Boxplots visualizing the distribution of the 14 untransformed features for 10 clusters, with the 10 features at the daily level displayed on the left and the 4 features at the weekly level on the right. The whiskers of the boxplots describe the [10, 90] percentiles.

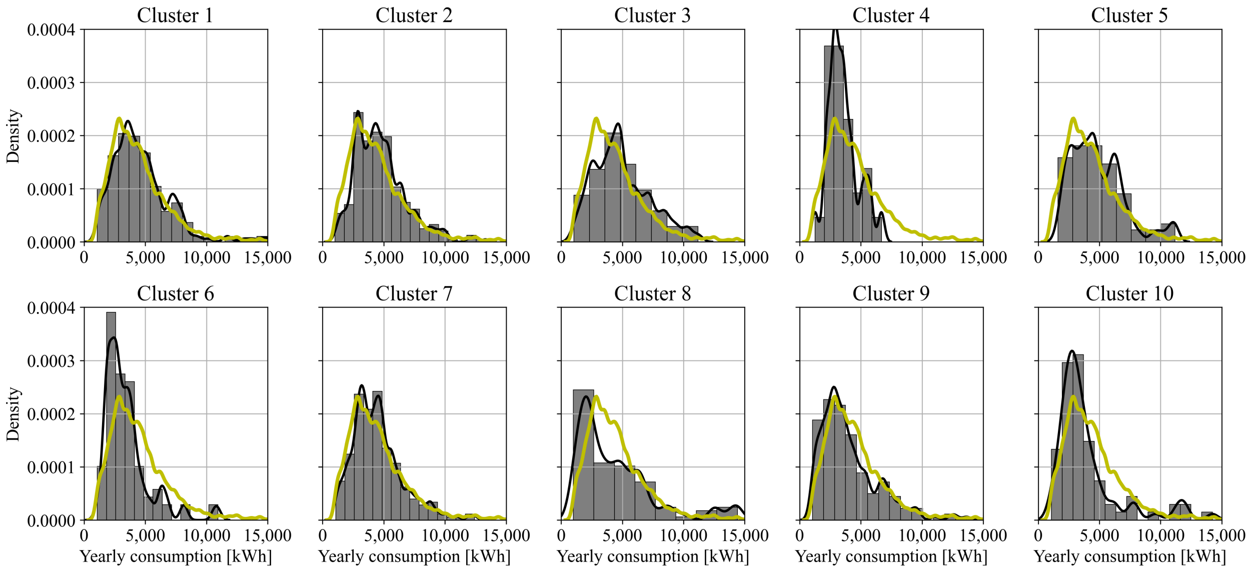

Figure 9. Density histograms with shared y-axes, displaying the distribution of the yearly consumption of the individual consumers assigned to each of the 10 considered clusters. The black plot denotes the density for the individual cluster, while the plot in yellow indicates the density of the full dataset.

It is clear that for each time period and in the same branch of the dendrogram, the differences between the fractions of total consumption for that period are limited. Rather, the temporal behavior of the peak demands is the driving force to further separate clusters in each of the three major branches of the dendrogram. Furthermore, the clustering process yields compact clusters with comprehensive results.

This illustrates the usefulness of a feature set that includes the temporal properties of peak demands, especially with the advent of capacity-based tariff schemes for low-voltage consumers. With the introduction of capacity-based tariffs, it is no longer sufficient to know when consumption occurs. Additional knowledge about when peak demands tend to happen is vital to offer consumers the most suitable techno-economic solution.

As a post-hoc validation of the performance of the proposed feature set in determining customer categories, the clusters of the different consumer types in the dataset as introduced in Section 2.1 are determined and given in Table 4. Clusters 4 and 5 are predominantely populated by households with electric heating, while cluster 10 groups households with high daytime consumption as well as the majority of the SMEs. However, not all profiles with electric heating are categorised inside clusters 4–5. This is further investigated in Figure 9, which displays the density plots of the yearly consumption for each individual cluster compared to the density plot of the full dataset. Matched against the density plot of the complete dataset, clusters 4, 6, 8 and 10 are skewed towards households with lower to average yearly consumption in the Eurostat classification. This distribution for cluster 4 is expected and can clarify the diffusion of households with electric heating over different clusters. As the demand profiles of these households can be considered an aggregation of the profile of a regular household with a load profile of an electric heating appliance, the features connected to the peak demands are intrinsically linked to the behavior of that load profile and the timing of the peak demands without the electric heating. The heating load profile for households with otherwise relatively low yearly consumption dominates the aggregated load profile, and consequently encounter the majority of their consumption peaks during the night, consistent with the behavior of cluster 4. For households with electric heating in e.g., cluster 3, the consumption and peak demands during the evening outweigh those during the night.

Table 4. Relative frequency of consumer categories over the 10 different clusters.

	# of Profiles	1	2	3	4	5	6	7	8	9	10
SME	20	0%	5%	0%	0%	0%	0%	0%	10%	10%	75%
Ripple control heating	107	2%	6%	6%	18%	36%	9%	20%	2%	2%	0%
Continuous heating	77	14%	8%	15%	11%	31%	1%	10%	4%	4%	0%
Air conditioning	18	17%	17%	6%	0%	6%	0%	33%	6%	17%	0%
Regular consumer	1206	17%	25%	5%	0.1%	1%	7%	17%	8%	16%	5%

It can be concluded that the proposed feature set is able to capture the known consumer categories from existing SLPs, and thus passes our self-imposed benchmark test. Three clusters can be attributed to known differences in behavior for low-voltage consumers: the presence of electric heating is captured in clusters 4–5, while the high daytime consumption of SME profiles is present in cluster 10. Deviations from these two clusters for electric heating can be traced back to differing contributions of the electric heating load to the total yearly consumption of the households.

3.2. Stochastic Nature of Peak Demands

The variability of the daily and weekly consumption and peak patterns are described by the entropy of their probability distribution, where the individual fractions are normalized with respect to the length of the considered time period. A uniform distribution with maximum uncertainty leads to a maximal value of the entropy, while the absence of uncertainty leads to an entropy value of 0.

For example, a situation where all peak demands occur during the night due to an electric heating would lead to 0 entropy at the daily level for the consumption probability distribution. The obtained distributions for the entropy at the daily and weekly level for the consumption and peak probability distributions of the full dataset are given in Figure 10. At the daily level, the peak demands exhibit a much larger variability than the consumption. This is unsurprising, given the continuous nature of the consumption. At the weekly level, this difference is less pronounced.

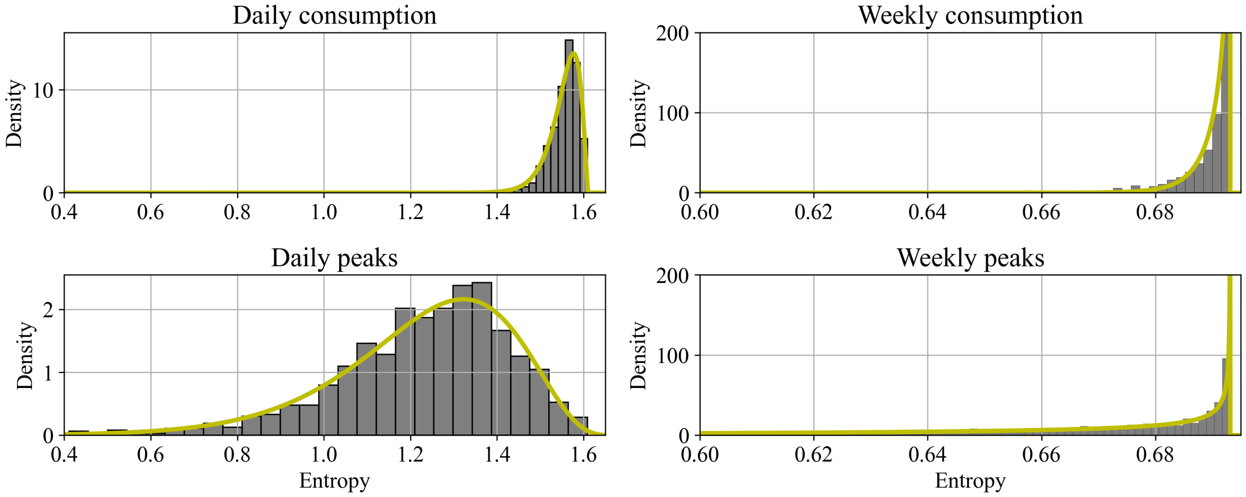

Figure 10. Density histograms of the entropy of the consumption and peak demands at the daily and weekly level, with a fitted beta probability density function overlaid in yellow.

A beta distribution was successfully fitted to each individual density histogram. The 2-parameter beta probability distribution, defined on the interval [0,1], is defined as follows, with $a > 0$ and $b > 0$:

$$f(x, a, b) = \frac{\Gamma(a, b)x^{a-1}(1 - x)^{b-1}}{\Gamma(a)\Gamma(b)} \tag{14}$$

The beta function offers several properties that make it suitable to describe the obtained distributions. First, it has a finite support: the regular 2-parameter beta function in Equation (14) has a [0,1] support. As the entropy can vary from 0 to a maximum of $-\ln(0.2)$ for the daily level and $-\ln(0.5)$ for the weekly level, the finite support of a rescaled and shifted beta function is appropriate. Second, as can be observed in Figure 10, the shapes of the daily and weekly behaviors differ significantly. The two shape parameters a and b in the definition of the beta probability function allow us to describe the four distributions with the same formula. For the distributions shown in Figure 10, it merely means that $b > 1$ for the distributions at the daily level, while $b < 1$ for those at the weekly level.

The relation between the entropy and the clusters obtained in Section 3.1 is investigated in Figure 11, which displays the mean values of the entropy for each individual cluster. The significantly lower entropy of the probability distribution describing the peak demands can be traced back to the clustering results. The overwhelming presence of peak demands during the night period results in low entropy for cluster 4, while cluster 10 exhibited a majority of its peaks during daytime. Similarly, half of the peak demands for cluster 1 occurred during the evening. On a weekly basis, clusters 6–7 showed a significant amount of peak demands during the weekend, leading to a lower entropy for this period. A low entropy of the probability function describing the peak demands can be taken as an indicator for the presence of a large amount of peaks in a certain time period, which can be leveraged to target demand response programs or peak shaving via an energy storage system. Furthermore, a clear relation can be observed between the obtained clusters on the introduced feature set and the entropy values of the peak demands. The lower entropy values in certain clusters can be traced back to differing intercluster consumer operations at the daily or weekly level.

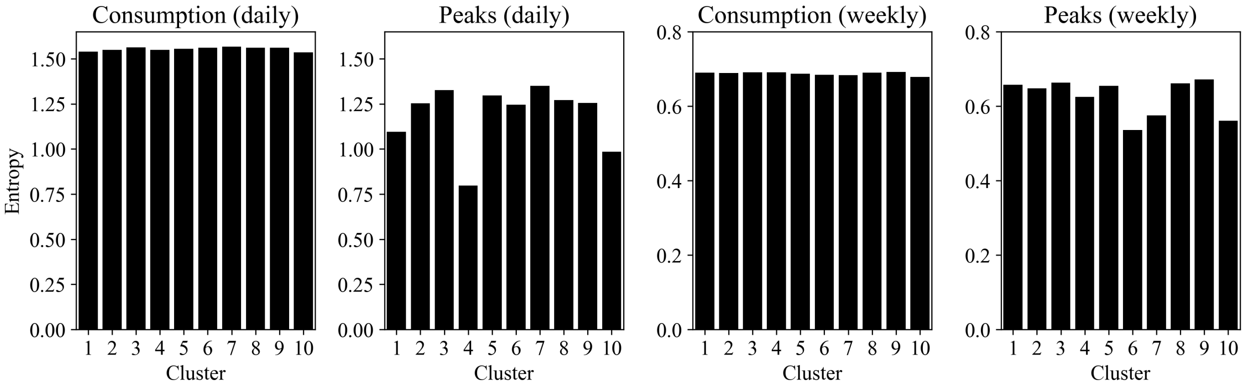

Figure 11. Mean values of the calculated entropy of the distributions of the normalized consumption and peak demands at the daily and weekly level of the individual consumers populating each cluster.

However, the stochastic nature of these peak demands remains an open question. The probability distributions of the peak demands tend to be significantly more variable than those of the consumption behavior, according to the entropy. Even so, this entropy as a single variable does not reveal anything about whether or not the amount of peaks in a certain time period is disproportional relative to the consumption in that time period.

Therefore, the Wasserstein-1 distance is used to quantify the difference between the probability distributions of the consumption and peak demands at the daily and weekly level for each individual consumer. A larger distance corresponds to a stronger deviation of the peak distribution from the distribution of the consumption, and thus peaks are more deterministic. Figures 12 and 13 yield the distributions for the Wasserstein-1 distances at the daily and weekly level, separated by individual cluster. Analogous to Figure 9, the distribution of the Wasserstein-1 distance calculated for each profile in the full dataset is included for comparison to cluster-specific behavior.

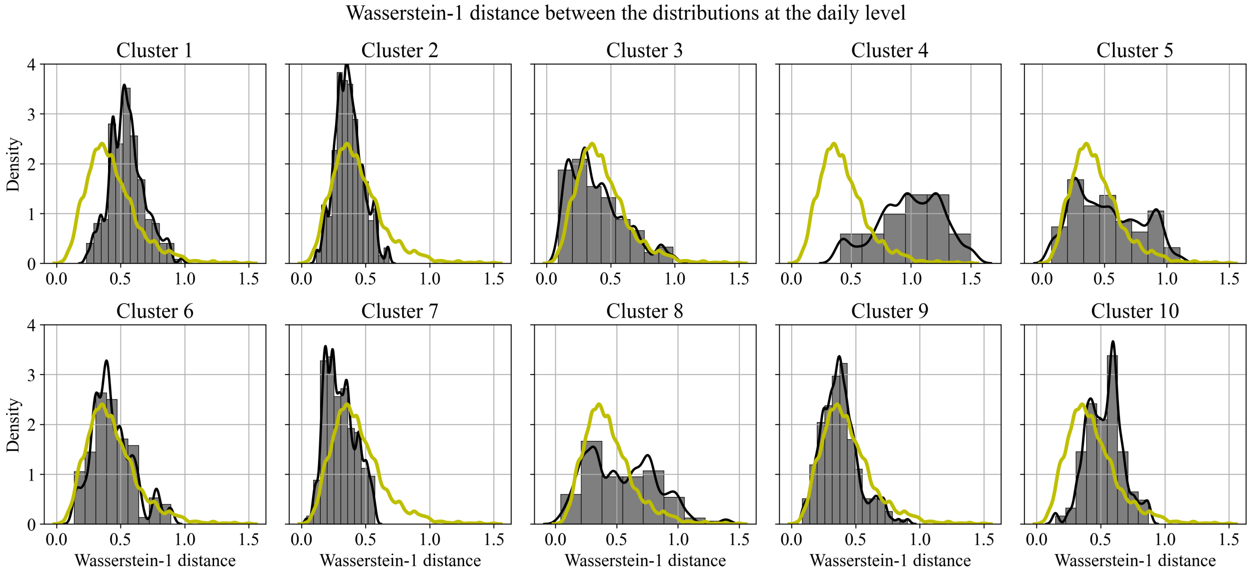

Figure 12. Histograms of the Wasserstein-1 distance between the distributions of the consumption and peak demands probability functions at the daily level. The black plot denotes the density for the individual cluster, while the plot in yellow indicates the density of the full dataset.

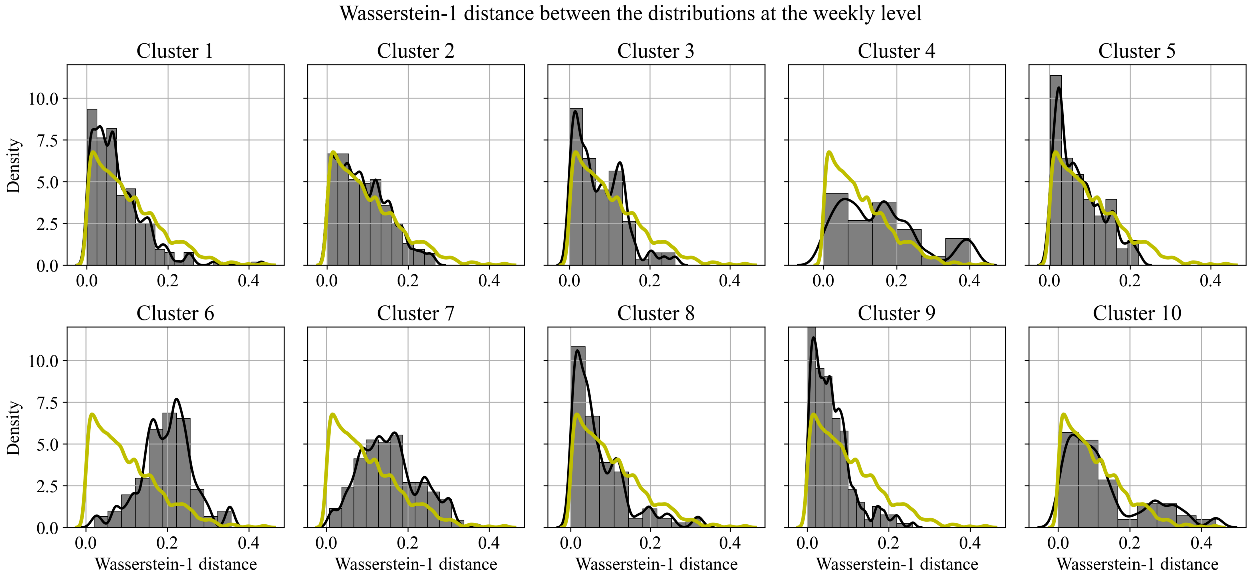

Figure 13. Histograms of the Wasserstein-1 distance between the distributions of the consumption and peak demands probability functions at the weekly level. The black plot denotes the density for the individual cluster, while the plot in yellow indicates the density of the full dataset.

The distributions of the Wasserstein-1 distances further confirm the findings concerning the behavior of consumers constituting each cluster. At the weekly level, clusters 6 and 7 show a major deviation from the dataset behavior, due to the presence of a disproportionate amount of peak demands in the weekend. Similarly for the daily level, cluster 4 displays a large Wasserstein-1 distance, pointing to the electric heating which pushes nearly all peak demands to nighttime.

Clusters 1 and 2 exhibited similar behavior for their consumption at the daily level in Figure 8. However, households in cluster 1 are characterized by an even higher amount of peak demands in the evening than those in cluster 2, translating to a higher than average Wasserstein-1 distance for cluster 1 at the daily level. This variability and disproportionate amount of peaks in a certain time interval offers insight in possibilities for targeted demand response initiatives or peak shaving via a residential energy storage system. While cluster 6–7 and 8–9 have a similar consumption pattern, the time of occurrence of peak demands is significantly different, which leads to distinct solutions.

As peak demands are typically generated by the simultaneous use of individual appliances, targeted demand response initiatives can be effective for cluster 6 and 7, where the majority of peaks occurs in the weekend. Scattering the use of individual appliances over different days or being mindful of the simultaneous use in the weekend by inducing behavioral changes can reduce the number of peak demands. However, this requires a trigger for the behavioral changes and for these appliances to be available in different time periods. If this is not an option, investing in an energy storage system applying a peak shaving algorithm during weekends, while e.g., maximising the PV self-consumption during weekdays could offer an alternative, although the economic viability depends on the local tariff structure and the investment cost. In contrast, cluster 8 is characterized by peak demands in the early morning and during the daytime, while households in cluster 9 exhibit peaks during the whole day. Consequently, for these households, a PV installation combined with a storage system can already offer a solution to reduce the demand from the grid, while maintaining a high self-consumption.

As a final check on the stochastic nature of peak demands, the relationship between the consumption in a time period and the presence of peak demands is investigated. Figure 14 displays the relations between the (untransformed) fractions of the consumption and peak demands at the daily level, with an ordinary least-squares (OLS) regression fit overlaid given the observed linear relation. The coefficients obtained in the OLS regression for $f^p = a \times f^c + b$, with f^p and f^c the fraction of respectively the peak demands and consumption in that time period, are given in Table 5. As the presence of electric heating heavily skewed previous results for the consumption and peak demands at night, consumers with and without electric heating are treated separately for this analysis.

A correlation between the fraction of the consumption and that of the peak demands is present in Figure 14 and 15. As the presence of consumption in a certain time period is a prerequisite for a peak demand, some relation between the two types of parameters was expected. At first sight, the linear relation could be interpreted as an indication of predictability of peak demands in a certain time period. However, it is the spread on this relation that is the indicator of the stochasticity of the peak demands. For example, if 30% of a household's total consumption is observed occurring during the evenings, the results shown in Figure 14 suggests that 30–60% of the peak demands can occur in this same time period. This large uncertainty, which is present for each of the considered time periods, severely limits the usability of this linear relation, observed for the full dataset.

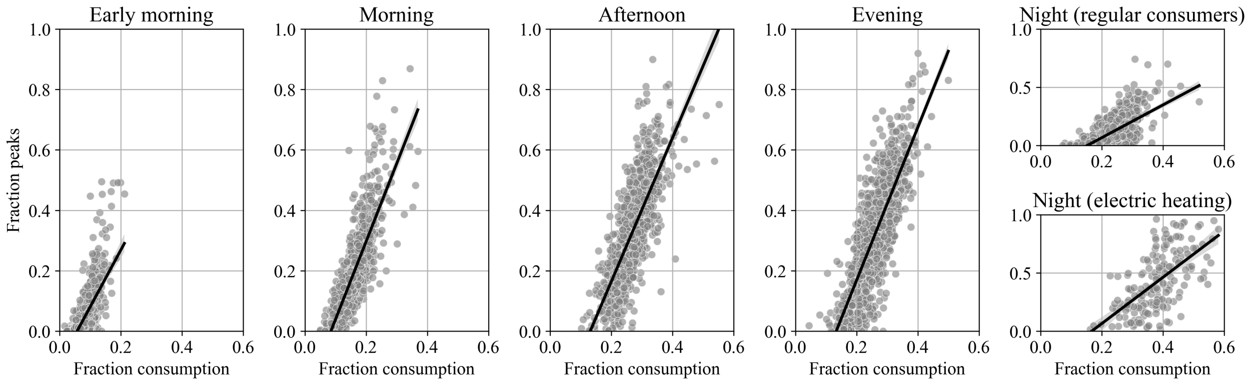

Figure 14. Relation between the fraction of consumption and peak demands in the time periods at the daily level, with an OLS regression estimate overlaid.

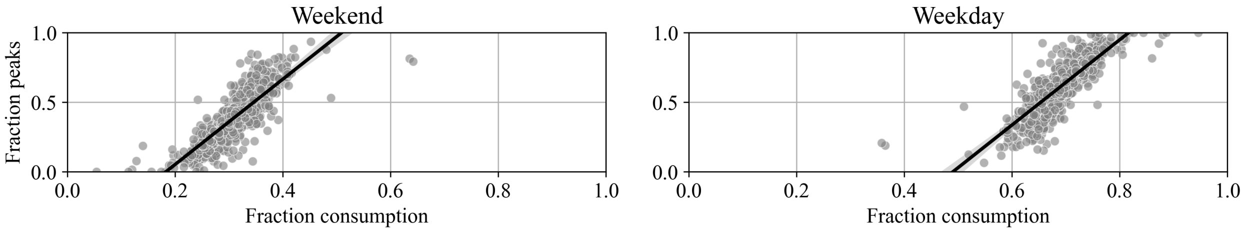

Figure 15. Relation between the fraction of consumption and peak demands in the time periods at the weekly level, with an OLS regression estimate overlaid.

However, the knowledge of the introduced clusters can partly alleviate this uncertainty. This is illustrated in Figure 16 for clusters 1–3, which group households with a large fraction of their consumption during the evening, with a high number of peak demands simultaneously occurring in this time period. While we should be cautious drawing conclusions based on clusters that only include a limited amount of households, it appears that the spread on the fraction of peak demands for the individual clusters is smaller than those in Figure 14 for the full dataset, while the linear correlation that was observed before is nearly non-existent in some relations.

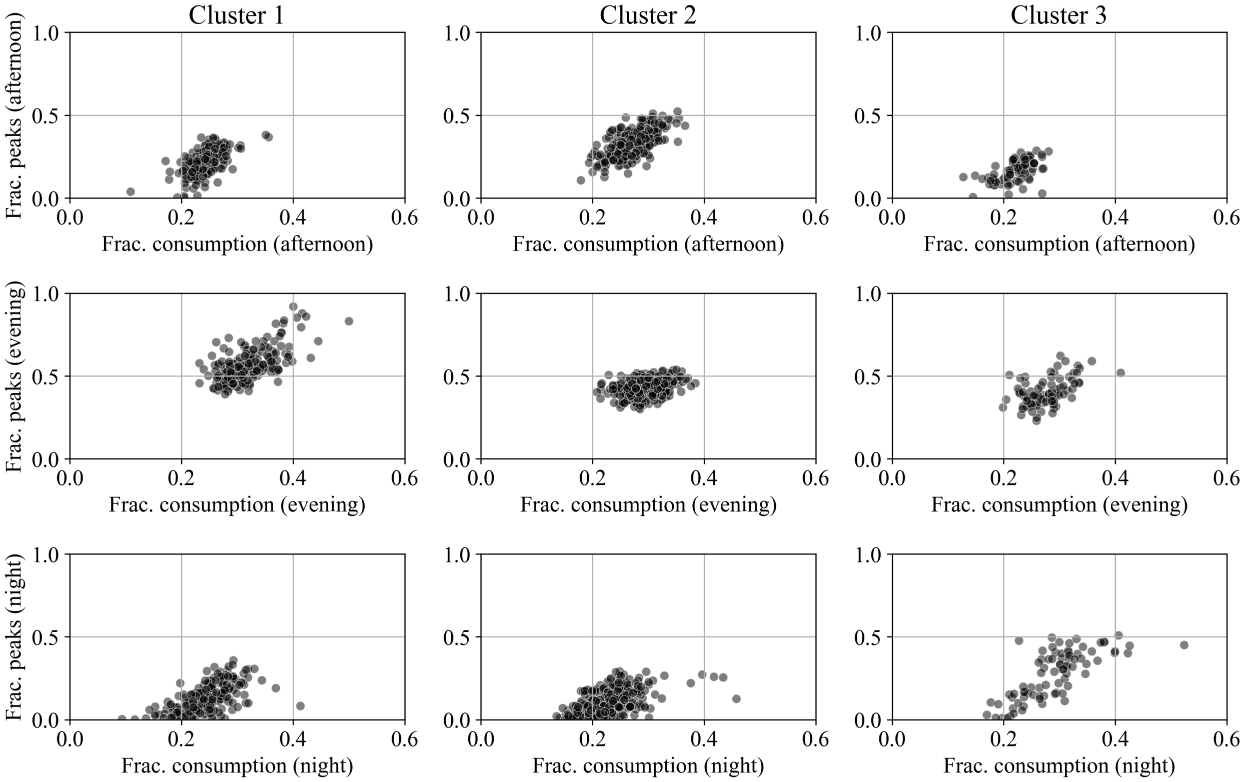

Figure 16. Relation between the fraction of consumption and peak demands in three time periods at the daily level for clusters 1–3.

Table 5. Coefficients of the OLS regressions shown in Figures 14 and 15 at the daily and weekly level.

Time Period	a	b
Early morning	1.88	−0.11
Morning	2.58	−0.22
Afternoon	2.34	−0.31
Evening	2.53	−0.33
Night (regular consumer)	1.41	−0.21
Night (electric heating)	1.99	−0.33
Weekday	3.06	−1.50
Weekend	3.06	−0.56

4. Conclusions

The introduction of capacity-based tariffs poses a major challenge for residential load modeling. The construction of representative load profiles traditionally involves an averaging process. However, this involuntarily leads to less volatile profiles and time-sensitive information about the peak demands is smoothened out. Consequently, we set out to construct a new feature set that would be able to capture the stochastic behavior of the peak demands. An expression for the load duration curve of individual low-voltage consumers was initially validated. Using the point of maximum curvature of the exponential decay as threshold, it was possible to define the individual's region of peak demands.

Two types of features were subsequently constructed. First, the fraction of consumption that occurs in a certain time period at the daily or weekly level. Second, the fractions of peak demands that occur in these same periods. The proposed feature set was used in a hierarchical clustering process to build 10 clusters from a dataset of 1.422 profiles of low-voltage consumers from a suburban region in Flanders. The clustering algorithm yielded compact clusters that showed a clear connection to real-life applications concerning the peak demands such as demand response initiatives, or the applicability of e.g., battery storage systems for peak shaving purposes.

Furthermore, differences in the behavior of the peak demands were found to be the main drivers of the clustering procedure. The presence of electric heating could be identified for several clusters, while others exhibited high daytime consumption during weekdays, which is typical for SMEs.

In the final analysis of this work, the stochastic nature of the peak demands was investigated by considering the relation between the consumption and the presence of peak demands in the same time period. The disproportionate presence of peak demands in a certain time period was quantified, and a linear relation was observed between the fraction of the consumption and peak demands in each time period. The spread on the results quantified the stochasticity of the peak demands, which limited the general applicability of the found relations. The obtained clusters showed a clear relation to the predictability and variability of the consumption and peak behavior, reducing the stochasticity of these peak demands and when they tend to occur.

The constructed feature set and performed clustering algorithm have several implications for the integration and application of new technologies on the low-voltage grid, which can be further investigated. First, knowledge about the time of occurrence of peak demands throughout the day and week allows for value stacking of residential energy storage systems combined with PV installations for capacity-based tariffs. This can be achieved by e.g., performing a control strategy that maximizes the PV self-consumption during the week, but applies a peak shaving algorithm during the weekend if peaks predominantly occur then. Second, the integration of electric vehicles for low-voltage consumers can be further investigated. Optimal charging schemes can be suggested based on periods where the household is typically occupied, but peak demands are absent. This could avoid additional costs related to higher peak demands for capacity-based tariffs. In both cases, the clustering result can be used to propose generic strategies for consumers with similar intracluster behavior. However, future experimental research and knowledge of upcoming capacity-based tariffs are both necessary to validate the economic and technical feasibility of these applications. A final possible research direction is related to the use of stochastic load models. The obtained relation between the fraction of consumption and peak demands in certain time periods can be used to validate and improve existing stochastic load models. This could allow for more accurate stochastic modeling of the temporal dependency of peak demands.

Author Contributions: Conceptualization, R.C. (Robbert Claeys) and J.D.; methodology, R.C. (Robbert Claeys), R.C. (Rémy Cleenwerck) and J.D.; software, R.C. (Robbert Claeys); validation, R.C. (Robbert Claeys), H.A. and J.D.; formal analysis, R.C. (Robbert Claeys); investigation, R.C. (Robbert Claeys); data curation, J.D.; writing—original draft preparation, R.C. (Robbert Claeys); writing—review and editing, R.C. (Robbert Claeys), H.A., R.C. (Rémy Cleenwerck), J.K. and J.D.; visualization, R.C. (Robbert Claeys) and R.C. (Rémy Cleenwerck); supervision, J.K. and J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: The authors would like to thank the Flemish distribution system operator Fluvius for providing the dataset of consumption profiles.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LDC	Load duration curve
OLS	Ordinary least-squares
PV	Photovoltaic
SLP	Synthetic load profile
SME	Small and medium-sized enterprises

References

1. VREG. Tariefmethodologie Voor Distributie Elektriciteit en Aardgas Gedurende de Reguleringsperiode 2021–2024. Available online: https://www.vreg.be/sites/default/files/Tariefmethodologie/2021-2024/BESL-2020-31/tariefmethodologie_reguleringsperiode_2021-2024.pdf (accessed on 10 November 2020).
2. NVE. Forslag til Endring i Forskrift om Kontroll av Nettvirksomhet. Available online: http://publikasjoner.nve.no/hoeringsdokument/2018/hoeringsdokument2018_08.pdf (accessed on 10 November 2020).
3. Thorvaldsen, K.E.; Bjarghov, S.; Farahmand, H. Representing Long-term Impact of Residential Building Energy Management using Stochastic Dynamic Programming. In Proceedings of the 2020 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Liege, Belgium, 18–21 August 2020; pp. 1–7.
4. Tureczek, A.M.; Nielsen, P.S. Structured literature review of electricity consumption classification using smart meter data. *Energies* **2017**, *10*, 584. [CrossRef]
5. Chicco, G. Overview and performance assessment of the clustering methods for electrical load pattern grouping. *Energy* **2012**, *42*, 68–80. [CrossRef]
6. Chicco, G.; Napoli, R.; Postolache, P.; Scutariu, M.; Toader, C. Customer characterization options for improving the tariff offer. *IEEE Trans. Power Syst.* **2003**, *18*, 381–387. [CrossRef]
7. Ozawa, A.; Furusato, R.; Yoshida, Y. Determining the relationship between a household's lifestyle and its electricity consumption in Japan by analyzing measured electric load profiles. *Energy Build.* **2016**, *119*, 200–210. [CrossRef]
8. Rhodes, J.D.; Cole, W.J.; Upshaw, C.R.; Edgar, T.F.; Webber, M.E. Clustering analysis of residential electricity demand profiles. *Appl. Energy* **2014**, *135*, 461–471. [CrossRef]
9. Kwac, J.; Flora, J.; Rajagopal, R. Household energy consumption segmentation using hourly data. *IEEE Trans. Smart Grid* **2014**, *5*, 420–430. [CrossRef]
10. McLoughlin, F.; Duffy, A.; Conlon, M. A clustering approach to domestic electricity load profile characterisation using smart metering data. *Appl. Energy* **2015**, *141*, 190–199. [CrossRef]
11. Yildiz, B.; Bilbao, J.; Dore, J.; Sproul, A. Recent advances in the analysis of residential electricity consumption and applications of smart meter data. *Appl. Energy* **2017**, *208*, 402–427. [CrossRef]

12. Bermingham, M.L.; Pong-Wong, R.; Spiliopoulou, A.; Hayward, C.; Rudan, I.; Campbell, H.; Wright, A.F.; Wilson, J.F.; Agakov, F.; Navarro, P.; et al. Application of high-dimensional feature selection: Evaluation for genomic prediction in man. *Sci. Rep.* **2015**, *5*, 10312. [CrossRef]
13. Al-Otaibi, R.; Jin, N.; Wilcox, T.; Flach, P. Feature construction and calibration for clustering daily load curves from smart-meter data. *IEEE Trans. Ind. Inform.* **2016**, *12*, 645–654. [CrossRef]
14. Verdú, S.V.; Garcia, M.O.; Senabre, C.; Marin, A.G.; Franco, F.J.G. Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Trans. Power Syst.* **2006**, *21*, 1672–1682. [CrossRef]
15. Manera, M.; Marzullo, A. Modelling the load curve of aggregate electricity consumption using principal components. *Environ. Model. Softw.* **2005**, *20*, 1389–1400. [CrossRef]
16. Jin, N.; Flach, P.; Wilcox, T.; Sellman, R.; Thumim, J.; Knobbe, A. Subgroup discovery in smart electricity meter data. *IEEE Trans. Ind. Inform.* **2014**, *10*, 1327–1336.
17. Haben, S.; Singleton, C.; Grindrod, P. Analysis and Clustering of Residential Customers Energy Behavioral Demand Using Smart Meter Data. *IEEE Trans. Smart Grid* **2016**, *7*, 136–144. [CrossRef]
18. Dent, I.; Aickelin, U.; Rodden, T.; Craig, T. Finding the creatures of habit; Clustering households based on their flexibility in using electricity. In *Clustering Households Based on Their Flexibility in Using Electricity*; 1 January 2012. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2828585 (accessed on 18 December 2020).
19. Räsänen, T.; Kolehmainen, M. Feature-based clustering for electricity use time series data. In *International Conference on Adaptive and Natural Computing Algorithms*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 401–412.
20. Masters, G.M. *Renewable and Efficient Electric Power Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2013.
21. Soyster, A.; Eynon, R. The conceptual basis of the electric utility sub-model of project independence evaluation system. *Appl. Math. Model.* **1979**, *3*, 242–248. [CrossRef]
22. Murphy, F.; Sen, S.; Soyster, A. Electric utility capacity expansion planning with uncertain load forecasts. *IIE Trans.* **1982**, *14*, 52–59. [CrossRef]
23. Poulin, A.; Dostie, M.; Fournier, M.; Sansregret, S. Load duration curve: A tool for technico-economic analysis of energy solutions. *Energy Build.* **2008**, *40*, 29–35. [CrossRef]
24. Motlagh, O.; Paevere, P.; Hong, T.S.; Grozev, G. Analysis of household electricity consumption behaviours: Impact of domestic electricity generation. *Appl. Math. Comput.* **2015**, *270*, 165 –178. [CrossRef]
25. Eurostat. Energy Statistics—Electricity Prices for Domestic and Industrial Consumers, Price Components. Available online: https://ec.europa.eu/eurostat/cache/metadata/en/nrg_pc_204_esms.htm (accessed on 10 November 2020).
26. Claeys, R.; Delerue, T.; Desmet, J. Assessing the influence of the aggregation level of residential consumers through load duration curves. In Proceedings of the 2019 IEEE PES Innovative Smart Grid Technologies Europe (ISGT-Europe), Bucharest, Romania, 29 September–2 October 2019; pp. 1–5.
27. Newville, M.; Stensitzki, T.; Allen, D.B.; Ingargiola, A. LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python. 2014. Available online: http://ascl.net/1606.014 (accessed on 18 December 2020).
28. Ward, J.H., Jr. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **1963**, *58*, 236–244. [CrossRef]
29. Virtanen, P.; Gommers, R.; Oliphant, T.E. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef]
30. Kang, J.; Lee, J.H. Electricity customer clustering following experts' principle for demand response applications. *Energies* **2015**, *8*, 12242–12265. [CrossRef]
31. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
32. Bolley, F. Separability and completeness for the Wasserstein distance. In *Séminaire de Probabilités XLI*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 371–377.
33. Ramdas, A.; Trillos, N.G.; Cuturi, M. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **2017**, *19*, 47. [CrossRef]