Experimental flexibility identification of aggregated residential thermal loads using behind-the-meter data

Ziras, Charalampos; Heinrich, Carsten; Pertl, Michael; Bindner, Henrik W.

Published in:
Applied Energy

Link to article, DOI:
10.1016/j.apenergy.2019.03.156

Publication date:
2019

Document Version
Early version, also known as pre-print

Experimental Flexibility Identification of Aggregated Residential Thermal Loads Using Behind-the-Meter Data

Charalampos Ziras[a,*], Carsten Heinrich[a], Michael Pertl[b], Henrik W. Bindner[a]

[a] *Center for Electric Power and Energy, Technical University of Denmark, Frederiksborgvej 399, Building 776, 4000 Roskilde, Denmark*
[b] *Austrian Power Grid, Wagramer Straße 19, 1220 Vienna, Austria*

Abstract

Thermal loads are an important source of flexibility at a residential customer level. The uncertain economic value of residential demand response (DR) and the rising customer data privacy concerns, require non-intrusive and economical approaches to harness flexibility. Baselines are essential for evaluating DR activations, however, the frequent use of flexibility makes them less accurate. In this paper, we first propose a baseline estimation method based solely on aggregated behind the meter data, which does not require additional knowledge of the portfolio's parameters. It is suited for frequent DR activations, and relies on a combination of linear interpolation, forward-backward autoregression and load decomposition. The method is then used to evaluate DR activations, in order to construct, and continuously update, a model for the response and the rebound behavior of the loads. A portfolio of 138 real residential customers equipped with electric heaters, and a large number of DR experiments, were used to verify the proposed approach. The response model, fitted with the experimental results, shows a strong dependency of the load reduction potential on time of day and ambient temperature, with a maximum load reduction equal to 1.2 kW per household. Validation results confirm that the fitted model can be used to estimate the response with a good accuracy. Finally, a model to describe and shape the rebound behavior of the loads is proposed and validated with real experiments.

Keywords: Aggregation Size; Baseline; Demand response; Experiments; Flexibility model; Rebound; Thermal loads.

*Corresponding author

Email addresses: chazi@elektro.dtu.dk (Charalampos Ziras), cahei@elektro.dtu.dk (Carsten Heinrich), michael.pertl@apg.at (Michael Pertl), hwbi@elektro.dtu.dk (Henrik W. Bindner)

Nomenclature

Abbreviations

BtM	Behind the meter
DER	Distributed energy resource
DR	Demand response
DSO	Distribution system operator
FBA	Forward-backward auto regression
ICT	Information and communication technology
IQR	Interquantile range
LI	Linear interpolation
LR	Linear regression
MA	Moving average
STD	Standard deviation
TSO	Transmission system operator

Parameters

β	LR coefficients
γ	duration of decomposition sub time series
λ	flexibility model coefficients
ϕ^{fw}, ϕ^{bw}	auto regression model parameters
$\varepsilon^{\mathrm{fw}}$, $\varepsilon^{\mathrm{bw}}$	auto regression model noise
ζ	number of weeks in the data set
a_i	share of released loads at step i
c^{fw}, c^{bw}	auto regression model constants
d_l	load reduction period for experiment l
n_{b}	number of experiments for rebound model
N_{tot}	length of time series
n_{tr}, n_{s}	load decomposition parameters
n_{ar}	auto regression model order
n_{tp}	auto regression training period
N_k	number of evaluation steps for test k
n_k	number of houses for test k
q	order of MA filter
t_k	starting time of test k
v	order of time representation in LR

Indices

j	baseline method index
k	test index
l	experiment index
t	time step

Variables

$\alpha_{j,t}$	linear combination coefficient for hybrid model
$\boldsymbol{X}$	regressors
$\boldsymbol{\epsilon}$	error of LR
$\hat{P}^{\mathrm{resp}}$	fitted average load response
$\hat{P}^{\mathrm{res}}$	estimated response
$\hat{w}_t$	detrended aggregated load
ω_l, ρ_l	experiment weights
$\overline{P}_i^{\mathrm{nor}}$	averaged normalized response at step i
e_l^{inst}	STD of instantaneous error per household at experiment l
e_l^{off}	STD of offset error per household at experiment l
e_t	instantaneous error value at step t
h	hour of the day
P_i^{sh}	normalized shaped load deviation at step i
P_l^{res}	average response at experiment l
$P_{l,i}^{\mathrm{nor}}$	normalized response during experiment l and step i
R_t	residual part of decomposition
S_t	seasonal part of decomposition
T^{amb}	ambient temperature
T_t	trend part of decomposition
u_t^j	estimated aggregated load of method j
u_t	estimated aggregated load
w_t	cleaned aggregated load
y_t	real aggregated load

1. Introduction

In 2017 Denmark set a new record, by generating 44% of its electricity demand from wind energy [1]. The total share of renewable energy in the electricity generation is expected to increase to approximately 87% in the next 10 years. Wind will remain the main renewable resource, and is expected to account for about two thirds of the renewable electricity generation [2]. This trend increases the uncertainty and variability of electricity production. The flexibility necessary to balance consumption and generation in the power system has historically been provided on the generation side. In recent years, the utilization of consumption flexibility, commonly referred to as demand response (DR), is becoming more popular.

Reduced costs for information and communication technology (ICT) infrastructure could make DR concepts, previously considered uneconomical, a viable solution. DR is expected to play a vital role in system balancing, through participation of distributed energy resources (DERs) in the wholesale market. In [3], power market structures and a large number of individual DR support mechanisms are reviewed. Since the available flexibility of individual customers is usually small, aggregators form and control pools of DERs. Through the combined control of many DERs, aggregators can reach bid sizes sufficiently high for the wholesale market. At the same time, DR can solve operational challenges of distribution system operators (DSOs). The increasing penetration of distributed generation, together with the ongoing electrification of heating and transport, will pose additional stress on distribution networks. To avoid expensive network retrofitting and upgrades, residential flexibility can be used to reduce equipment loading during hours of peak consumption or peak generation [4].

A number of works have investigated how local DSO markets that incorporate DR could be designed. In [5], a flexibility clearing house is proposed in parallel to the wholesale markets, where aggregators can place offers which help relieve congestions in the local network. In [6], the concept of a proactive distribution company is introduced, which is active on the wholesale markets, and at the same time operates a local distribution-level market, to increase network efficiency and renewable energy integration. The work of [7] proposes a local market structure for distribution services, where individual prosumers directly offer their flexibility to the local DSO. In [8] a market structure where capacity can be reserved on a weekly basis, and for each hour of the day, is proposed. In [9], the authors propose a vertically integrated market for flexibility services, where DSOs and the Transmission System Operator (TSO) can procure DR services simultaneously.

Recognizing the increasing importance of DR utilization on a distribution level, the Danish project EcoGrid 2.0 develops, implements and tests a market approach for DR services both on a TSO and DSO level. In other words, aggregators can offer services to the DSO while participating in the wholesale markets. The goal of this paper is to characterize the flexibility of residential thermal loads under practical limitations, such as the lack of dedicated metering (behind the meter (BtM) data), uncertain customer behavior, and zero information regarding the loads' characteristics. An additional challenge arises from the need to describe and utilize the flexibility of a relatively small number of loads. This is necessary for DSO service provision on low aggregation levels, such as medium or low voltage feeders.

In subsection 1.1 different flexibility modelling approaches and the advantages of the proposed experimental identification of flexibility are discussed. In subsection 1.2 the idea of baselines is introduced, and various proposed methods to estimate baseline consumption are discussed. In subsection 1.3 the contributions and the organization of the paper are presented.

1.1. Flexibility modelling under limited information

There is extensive literature on thermal loads modelling and flexibility aggregation. Many works related to DR from thermal loads focus on modelling larger buildings. For example, [10] proposes a model predictive approach for the control of a building's thermal system, participating in a demand response program. Due to the complicated resulting optimization problem, a heuristic procedure is proposed to reduce the computational time.

A number of works have also focused on identifying the flexibility of the thermal systems of small residential customers. The potential of two different residential heating systems to provide flexibility is assessed in [11], assuming knowledge over a large number of building characteristics. The authors of [12] do not rely on very detailed building information, and propose to numerically fit a 2R2C model for each room of a household. This approach requires air temperature and space heating power data for each room. There is a number of challenges associated with such a detailed modelling approach. First, room air temperature data with good quality may not be available. Second, under BtM metering the heating power of each room is also not directly available. Third, the layout of the house affects temperature dynamics. Fourth, a variety of factors such as the windows surface, the house orientation, the effect of furniture on the thermal mass [13], imperfect weather data, and user disturbances (window/door openings, occupancy, cooking, etc.), make system identification a challenging and time-consuming task.

Apart from the difficulty of identifying and validating the model of each household separately, it is hard to quantify and assess the impact of the various sources of uncertainty on smaller aggregations of residential buildings. Furthermore, it may also be difficult to aggregate such complicated models, and express the flexibility of a population of thermal loads. For these reasons simpler and more generic models to quantify flexibility have been proposed. In [14], a battery model is used to characterize the flexibility of an aggregation of small thermal loads, but assuming perfect parameter knowledge. A similar model is proposed in [15], together with a practical, experimental-based method to identify the battery model's parameters. However, this method assumes dedicated metering and neglects the various user-induced disturbances. Other works using a similar approach, such as [16], focus on aggregating the flexibility of large commercial buildings with more complex dynamics, where knowledge of the individual load's parameters is necessary.

In [17] a generic, bottom-up way of expressing flexibility is proposed, similar to the computationally demanding, geometric approach using polytopes [18]. The work of [19] also follows a bottom-up approach in line with [17], focusing on building energy systems. The authors of [20] introduce a flexibility index, and the authors of [21] an instantaneous power flexibility indicator to quantify the flexibility of the thermal systems of buildings. However,

these approaches are suited for larger loads, and require high knowledge of the underlying flexible units. Therefore, they are not easily scalable.

Apart from the aforementioned difficulties in performing individual system identification, expressing uncertainties, and aggregating flexibility in residential thermal loads, other data-related problems arise in a real-world implementation. Bottom-up and data-intensive approaches require intrusive user profiling to model occupancy and the customers' actions. This raises important data-privacy concerns, whereas the customers' engagement and consent to the use of their private data is not certain. Additionally, the aggregators' access to data generated by smart appliances, such as thermostats, is not guaranteed. For example, it may not be straightforward for an aggregator to use historical data generated by smart thermostats, due to customer agreements with the devices' manufacturer, or a previous aggregator who was providing services to the customer. Therefore, even if user privacy issues are neglected, the required data may not be accessible.

Given the high uncertainty regarding the economic value of residential DR [22], a framework which aims at reduced costs is investigated in this work. To this end, a simple centralized control setup with limited information is used. The benefits of centrally coordinated actions under network constraints, a situation similar to a DSO service for load reduction, were highlighted in [23]. The proposed approach overcomes the aforementioned drawbacks, as it relies on the use of aggregated meter data, and the evaluation of DR activations. Customers also equipped with PV units and battery systems may have more advanced controls and a higher level of information. Such cases are outside the scope of this work. The basic principle of our approach is that an aggregator can continuously use the evaluation results of DR activations as training data to update the flexibility model. Characterizing flexibility by evaluating DR activations allows this continuous re-training process, while the aggregator uses the portfolio for spot-price optimization, offering balancing services, etc. A minor drawback of such an approach is that in the initial stage larger mismatches between the expected and actual flexibility provision will appear. However, the continuously collected evaluation data can be used by the aggregator to update the flexibility model, and achieve more accurate estimations of its flexibility provision.

1.2. Baselines

To evaluate the performance of the households during the experiments, it is required to estimate their consumption in the case where the experiment had not taken place. This hypothetical consumption is called a baseline. The baseline concept can be used for two fundamentally different purposes. First, baselines are used for defining and verifying DR services [22]. In this case, the baselines must be defined in a fair way, either by the service buyer or an independent third party. Second, baselines can be used by an aggregator to assess the impact of its control actions on the performance of its portfolio. The purpose of these two baselines may seem similar: to quantify the flexibility provision of an aggregation of customers under a DR activation. The first type of baselines is calculated with an agreed upon methodology, but can be manipulated by the flexibility providers to increase their profits [24]. In this paper we use baselines to characterize the flexibility of thermal loads,

and in this case the baseline must reflect the actual estimated uncontrolled behavior of the loads, absent of any gaming behavior.

Many methods have been proposed to construct baselines. In [25], five different baseline models are investigated. These models are based on linear regression (LR) of varying complexity, using ambient temperature, weighted average energy use, time of day and day of week. The performance of baseline models considering a linear combination of hourly consumption and temperature data was also investigated in [26] and [27]. A weather-adjusted variant of these models was proposed in [28], and adding occupancy data was shown to improve performance in [29]. A problem of these methods is that they may create a significant offset when producing a baseline estimation for a short period of time. A simple way to overcome the offset problem is to use linear interpolation (LI) between the power consumption values before and after a DR event. Such a method to produce baselines was applied in [30] and in [31], in order to quantify the flexibility of large thermal loads under DR activation.

The authors of [32] use the consumption of a reference population to create a baseline for a population of loads participating in a DR experiment. The consumption of the reference population for four hours preceding the DR activation is scaled, by using least-squares fitting, to match the consumption of the population that is participating in the DR activation. Then, a smoothing spline is fitted to create the baseline. A similar approach, using the consumption of non-DR participants to estimate the baseline of DR participants is proposed in [33] and in [34]. The problem with this practical approach is that always a reference population with similar characteristics needs to exist, which does not participate in DR activations. Therefore, a baseline cannot be created for the whole population. Furthermore, this approach becomes problematic under frequent DR activations. The authors of [35] use an adjusted autoregressive model for short-term electricity forecasting, where the proposed method outperformed a neural network model. Support vector regression was used in [36] to calculate baseline for commercial buildings using occupancy and weather data as regressors.

1.3. Contribution and paper organization

Two are the main contributions of this paper:

- A scalable, non-intrusive, data-driven methodology for identifying and modelling the flexibility of residential thermal loads by conducting a series of DR experiments is proposed. As more DR experiments are carried out over time the flexibility model accuracy can be continuously improved. This methodology considers only aggregate BtM energy data and average weather data, whereas it requires zero knowledge of individual household parameters. Additionally, the derived flexibility model is validated with real experiments involving a large number of thermal loads.

- A method to construct baselines, which are used to evaluate the results of the conducted experiments, is proposed. This method considers a large number of periods of DR activations, requires only aggregate BtM and weather data, and is evaluated on real households metering data. The method accounts for a large numbers of periods

of DR activations, by first removing their effect on the aggregated consumption of the loads. Next, a linear combination of three methods, namely LI, forward-backward autoregression (FBA), and load decomposition, is used to create baselines.

The remainder of the paper is organized as follows. In Section 2, the EcoGrid 2.0 project and the conducted experiments are described. In Section 3, the different investigated baseline methodologies are described, whereas in Section 4 the performance of these methods is evaluated. In Section 5, a flexibility model, based on the analysis of a large number of conducted experiments, is presented and validated. Finally, Section 6 concludes the paper.

2. The EcoGrid 2.0 project

In this section an overview of the EcoGrid 2.0 project is given. More specifically, in subsection 2.1 background information regarding the control setup, the metering infrastructure, and the purpose of the project is provided. In subsection 2.2 the proposed EcoGrid 2.0 market setup is presented, and in subsection 2.3 the DR experiments conducted in the course of the project are described.

2.1. Background

EcoGrid 2.0 is a demonstration project which investigates how the flexible consumption of residential customers can be utilized for offering power system services at a TSO and DSO level. The customers participating in the project are located in the Danish island of Bornholm [37]. It is the continuation of the EcoGrid EU project, where the use of real-time 5 minute price signals to shift consumption, and thus balance the power system, was studied [38]. In the EcoGrid EU project, residential customers were equipped with smart meters and communication/control infrastructure to participate in DR experiments. These smart meters have the capability of metering and storing active and reactive power consumption in 5 minute intervals, instead of the typical 15 or 60 minutes of common meters. These values are available in a central database (called DataHub) with a delay of $12 - 36$ hours. An overview of the EcoGrid 2.0 project setup can be seen in Fig. 1, where aggregators participate in both TSO and DSO markets.

The customers' flexible load consists of electric heaters and heat pumps, with roughly half of the customers covering their heat demand with each type of heating load. The electric heaters are controlled by adjusting room temperature setpoints. In the case of heat pumps, a throttle signal can be sent, which prohibits them from switching on. More details on the control of those heat pumps can be found in [32]. The flexible load is metered together with the rest of the household consumption, as well as photovoltaic production, if it exists.

An implication of the EcoGrid 2.0 setup is that online observations of the loads consumption are not possible, due to the significant data transmission delay. This means that if an aggregator wants to perform closed-loop control, an upgrade of the measurement capabilities is needed, either by improved transmission time or by separate measurement infrastructure. Since meters are owned by the DSO, there are restrictions on how they can be adjusted to

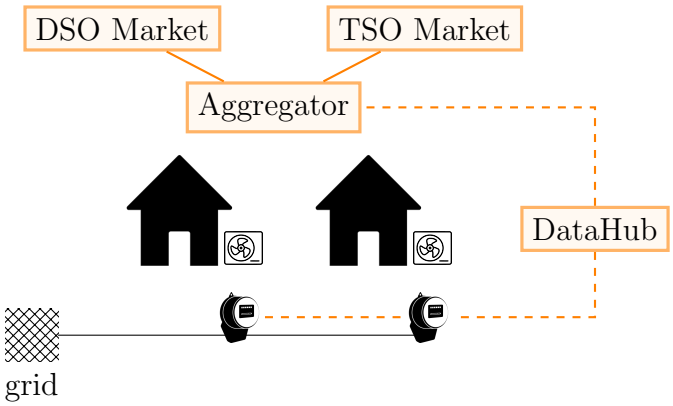

Figure 1: Overview of the EcoGrid 2.0 setup.

accommodate advanced capabilities. It is also probable that customers may change aggregators in the future, which could make the installation of additional meters uneconomical for small consumers.

The methods developed in EcoGrid 2.0 are based on data which will be accessible in Denmark in the near future. By the end of 2019, all residential customers across Denmark will be equipped with smart meters and their power consumption data will be centrally stored in the DataHub. Identifying flexibility only via the readily available smart data has the advantage that aggregators do not need to install additional metering equipment, and can also use historical data. If manufacturers equip DERs in the near future with ICT capabilities, which is already the case for many electric vehicles chargers, aggregators will be able to utilize DER flexibility without installing additional ICT infrastructure. By having an accurate representation of flexibility, the aggregators can offer balancing power to the TSO, or fulfill possible commitments to the DSO.

2.2. EcoGrid 2.0 market

In EcoGrid 2.0 an asymmetric balancing market with a 15 minute granularity is proposed. This market is based on the work of [39], where asymmetric block offers are included in the economic dispatch of balancing power. This is done to facilitate the participation of energy-constrained units in the balancing market. An example of such units are heating loads, which can only shift their consumption. Therefore, a load change (referred to as response) is followed by a predictable load change in the opposite direction (rebound). The rationale behind the asymmetric balancing market is that including the rebound in the dispatch reduces the overall balancing costs. The TSO has knowledge of the imbalance which will be caused by the rebound, and the aggregator is not subject to the balancing price uncertainty for such an event. As shown in [39], such a market setup can reduce system balancing costs.

Apart from a modified balancing market, a DSO market is also proposed. In this market the DSO buys flexibility services in specified connection points of the distribution grid to alleviate potential operational issues. Two DSO services are being investigated in the

course of the project. One is load changes around a baseline consumption. In this service, the aggregator agrees to reduce its load by the agreed amount and for a specified time period, while maintaining any subsequent load increase below a contracted limit, always with reference to the baseline profile. The second DSO service is capacity limitation, where the aggregators agree to cap their consumption to a certain level for a specified duration.

2.3. Demand response experiments

A number of experiments were conducted to identify flexibility, as well as model and shape the rebound consumption. Control was performed by changes in the thermostat setpoints. To switch heaters off, the thermostat setpoints were decreased to a very small value, which is equivalent to directly switching the heating loads off. Lowering the setpoints causes a reduction of consumption, followed by an increase when the setpoints are reset to their original values. The data used in this paper covers approximately 6 months, beginning from the 1st of September 2017 and corresponds to 138 loads equipped with electric heaters. The periods of load reduction lasted $30 - 60$ minutes to avoid user discomfort. The experiments were conducted under varying ambient temperature and time of day conditions, to assess the impact of these factors on the aggregation's flexibility.

3. Baseline Methodology

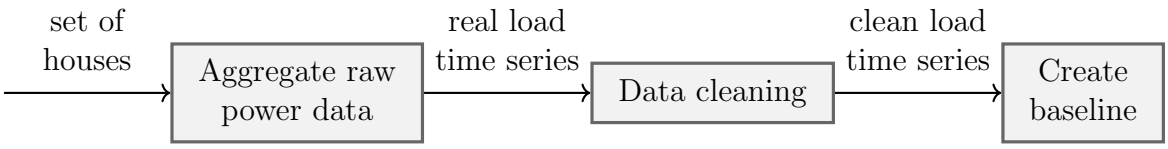

Figure 2: Schematic overview of the evaluation process of the experiments.

The steps of the baseline calculation process are depicted in Fig. 2. The first step is to define the set of houses for which the baseline will be created. Afterwards, the meter data is summed up to get the real aggregated load of the population. Due to the frequent control of the flexible load by the aggregator, this aggregated load time series contains large deviations from the "typical" consumption patterns. These deviations are not necessarily related to the weather, time of day or user behavior, but to the aggregator's control actions. An example of deviations caused by DR experiments can be seen in Fig. 3. These DR activations were part of the conducted experiments; in a commercial application setup, such activations can be more frequent. The goal of the baseline model is to provide an estimation of the natural consumption of the aggregation. Therefore, before using the time series as training data for the baseline model, the influence of external control must be removed. In subsection 3.2 we elaborate on the data replacement process, where a *clean* aggregated load time series is derived from the *real* one. This clean load time series is used as training data to create the baseline consumption estimation for each evaluation period. The investigated methods are described in subsection 3.3. Before describing the data replacement and baseline creation

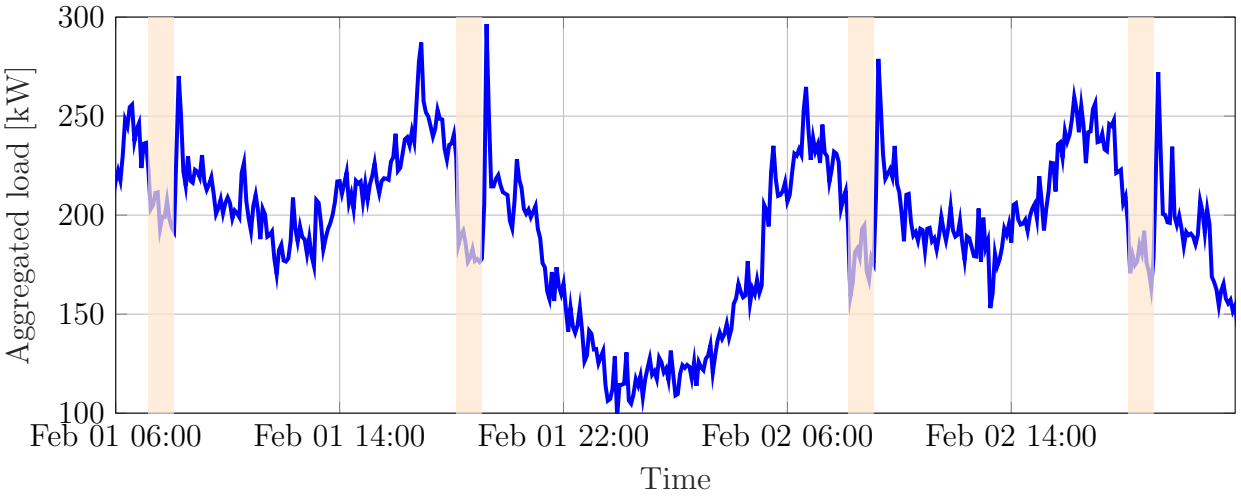

Figure 3: Aggregated load for all households under frequent DR activations. The load reduction periods are marked with orange color.

processes, the error metric used in the evaluation of the different methods is introduced in subsection 3.1.

3.1. Error metric

We denote the real aggregated load of a collection of residential customers by y_t at each time step t, and by u_t the estimated load, given by one of the different investigated methods. The deviation of the estimated load from the actual load (expressed in kW) is used as an error metric. For each $k-$th evaluation period of N_k steps and starting at step t_k, the instantaneous error values are calculated as

$$e_i^k = u_{t_k+i}^k - y_{t_k+i}, \quad \forall\ i \in \{1, ..., N_k\}. \tag{1}$$

To simplify notation, bold letters will represent vectors and matrices throughout the paper. Additionally, l will be used to index the DR experiments, and k will be used to index tests used for evaluating the various baseline methods. Different metrics, such as the standard deviation (STD) or percentiles of the error distributions, can be used to assess the performance of the different baseline methods. Due to the metering data granularity of 5 minutes, the aggregated load is relatively noisy. The high-frequency load fluctuations are very hard to estimate, if not impossible in practice. To avoid the high frequency fluctuations, a moving average (MA) filter is applied to $\boldsymbol{y}$, to obtain the smoother aggregate load $\tilde{y}_t^q$

$$\tilde{y}_t^q = \frac{\sum_{j=1}^{q} y_{t-(q+1)/2+j}}{q}, \quad \forall\ t \in \{(q+1)/2, ..., N_{\text{tot}}\}, \tag{2}$$

where N_{tot} is the length of time series $\boldsymbol{y}$ and q is the order of the MA filter. The corresponding difference is calculated as

$$\tilde{e}_t^q = \tilde{y}_t^q - y_t, \quad \forall\ t \in \{(q+1)/2, ..., N_{\text{tot}}\}. \tag{3}$$

This error value represents the error due to high-frequency oscillations, which we believe are unpredictable. Hence, $\tilde{e}^q$ is used throughout this paper as a benchmark, to provide lower bounds for the errors. The MA order was set equal to $q = 5$, resulting in the smoother aggregated load profile shown in Fig. 4. Notice the sudden load variations occurring with a frequency of 5 minutes, which are very hard to estimate.

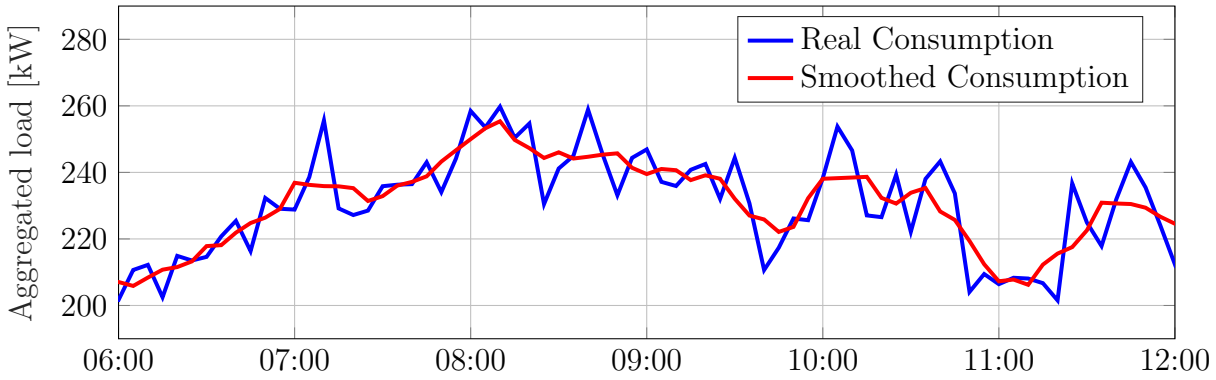

Figure 4: Actual and smoothed aggregate load for a morning of a typical winter day.

3.2. Data replacement during experiments

The need to obtain a clean aggregated load time series (denoted by $\boldsymbol{w}$) has already been described. The data replacement methods can only be evaluated during periods unaffected by the DR experiments. The reason for this is that during the experiments, or the subsequent rebound periods, the aggregated load is significantly affected by the aggregator's control actions. Therefore, the accuracy of the baseline estimation methods cannot be validated. A large number of DR experiments was conducted, with durations ranging between 30 and 60 minutes. Any load rebound effects after 1 hour and 2 hours respectively were negligible. The periods when experiments were conducted were excluded from the evaluation of the data replacement, along with a succeeding rebound period equal to two times the experiment duration. LI and FBA were evaluated as data replacement methods on randomly chosen periods which do not overlap with the DR experiments.

3.2.1. Linear interpolation

Linear interpolation is the most straightforward data replacement method. The interpolated aggregated load values are calculated as

$$u_t^k = y_{t_k} + \frac{y_{t_k+N_k+1} - y_{t_k}}{N_k}\left(t - t_k\right), \; t \in \{t_k + 1, ..., t_k + N_k\}. \tag{4}$$

3.2.2. Forward-backward auto regression

For each test period, two auto regressive models of order n_{AR} are trained, once for forward data and once for backward data. The forward model expresses the value y_t as a

linear combination of previous values plus an error term, and can be expressed as

$$y_t = c^{\text{fw}} + \sum_{i=1}^{n_{\text{AR}}} \phi_i^{\text{fw}} y_{t-i} + \varepsilon_t^{\text{fw}}. \tag{5}$$

The reverse model expresses the value y_t as a linear combination of future values. It can be expressed as

$$y_t = c^{\text{bw}} + \sum_{i=1}^{n_{\text{AR}}} \phi_i^{\text{bw}} y_{t+i} + \varepsilon_t^{\text{bw}}. \tag{6}$$

ϕ_i^{fw} and ϕ_i^{bw} are the linear model coefficients, c^{fw} and c^{bw} are constants, and $\varepsilon_t^{\text{fw}}$ and $\varepsilon_t^{\text{bw}}$ represent the error terms at time t. The forward model is trained on n_{TR} time steps before t_k, and the backward model n_{TR} time steps after t_{k+N_k}. Once both models are trained, (5) and (6) can be used to forecast (or backcast) the most likely values for the missing time steps recursively, using the fact that the expected value of ε is equal to zero. Finally, the two results are added up by weighing each result by the distance to the next known value. It is important to clarify that in order to evaluate the FBA method, it must be first applied on all periods where DR experiments were conducted. Once the influence of the experiments on the aggregated load is removed, FBA is then applied on each evaluation period individually.

3.2.3. Data replacement results

We compare the performance of the two methods on 300 randomly chosen 3-hour periods ($N_k = 36$ for all periods). For the FBA model, a horizon of 2 days was chosen, and a training period of 18 days ($n_{\text{ar}} = 576$ and $n_{\text{tp}} = 5184$). These parameters were optimized by performing an exhaustive search on n_{ar} and n_{tp} values. FBA is first applied to remove the influence of all DR experiments, and then applied for each test separately, as described earlier. In the case of LI, there is no need for data replacement during the DR experiments, since the evaluation periods never coincide with the DR experiments. The performance of the two methods is shown in Fig. 5. The STD and the interquantile range (IQR) are used as metrics. The subscript of IQR denotes its range, i.e., a value of 95 indicates the difference between the 97.5^{th} and the 2.5^{th} percentiles. FBA outperforms LI in all metrics, and it will be used as a data replacement method for the rest of the analysis in the paper. More advanced methods based on load decomposition, which are presented later, cannot be used for replacing data for a large number of periods concurrently.

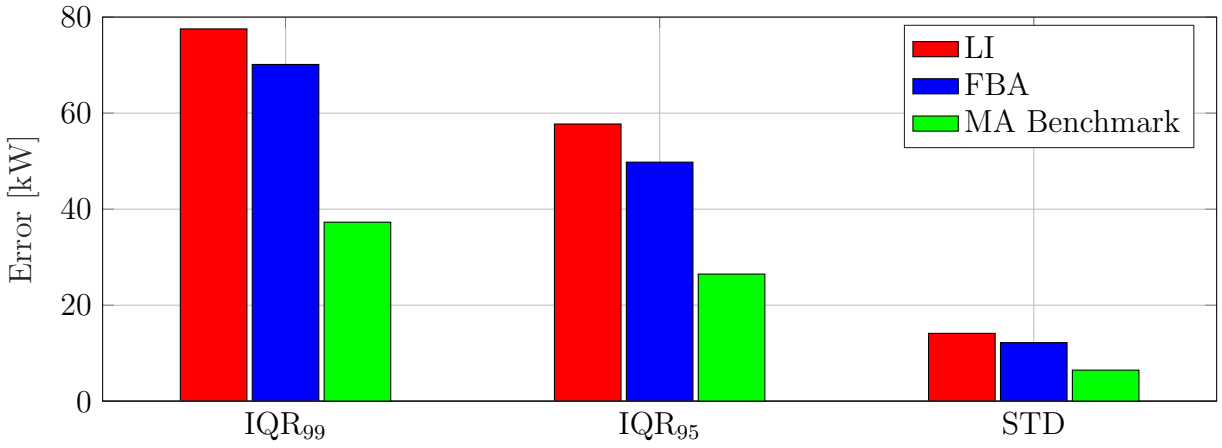

Figure 5: Comparison of LI and FBA methods for data replacement, against the benchmark case.

3.3. Baseline methods

After obtaining a clean time series $\boldsymbol{w}$ to use as training data, different baseline methods can be applied. An overview of these methods is given in Table 1. The first two methods, LI and FBA, were introduced in the previous subsection. We further investigated three methods based on a decomposition technique, which is described later in more detail. In the first, the residual load component obtained from the decomposition is neglected. In the second, FBA is applied on the residual. In the last variant, LR is applied on the residual. Finally, a hybrid approach, based on a linear combination of the aforementioned five baseline methods is presented.

Method	Description
M1	Linear Interpolation
M2	Forward-Backward Auto regression
M3	Decomposition + Ignoring Residual
M4	Decomposition + FBA on Residual
M5	Decomposition + Linear Regression on Residual
M6	Hybrid

Table 1: Overview of the different baseline methods.

3.3.1. Load decomposition

We apply a decomposition technique from [40], which is based on locally weighted scatter plot smoothing, on the clean aggregated load time series $\boldsymbol{w}$. The benefits of deseasonalization and detrending with respect to forecasting have been shown in [41, 42]. This subsection outlines the concept of the applied decomposition method; for further details, the reader is referred to [40]. The method splits time series $\boldsymbol{w}$ into a trend component $\boldsymbol{T}$, a seasonal component $\boldsymbol{S}$ and a remainder component $\boldsymbol{R}$, such that

$$w_t = T_t + S_t + R_t, \quad \forall\ t \in \{1, ..., N_{\text{tot}}\}. \tag{7}$$

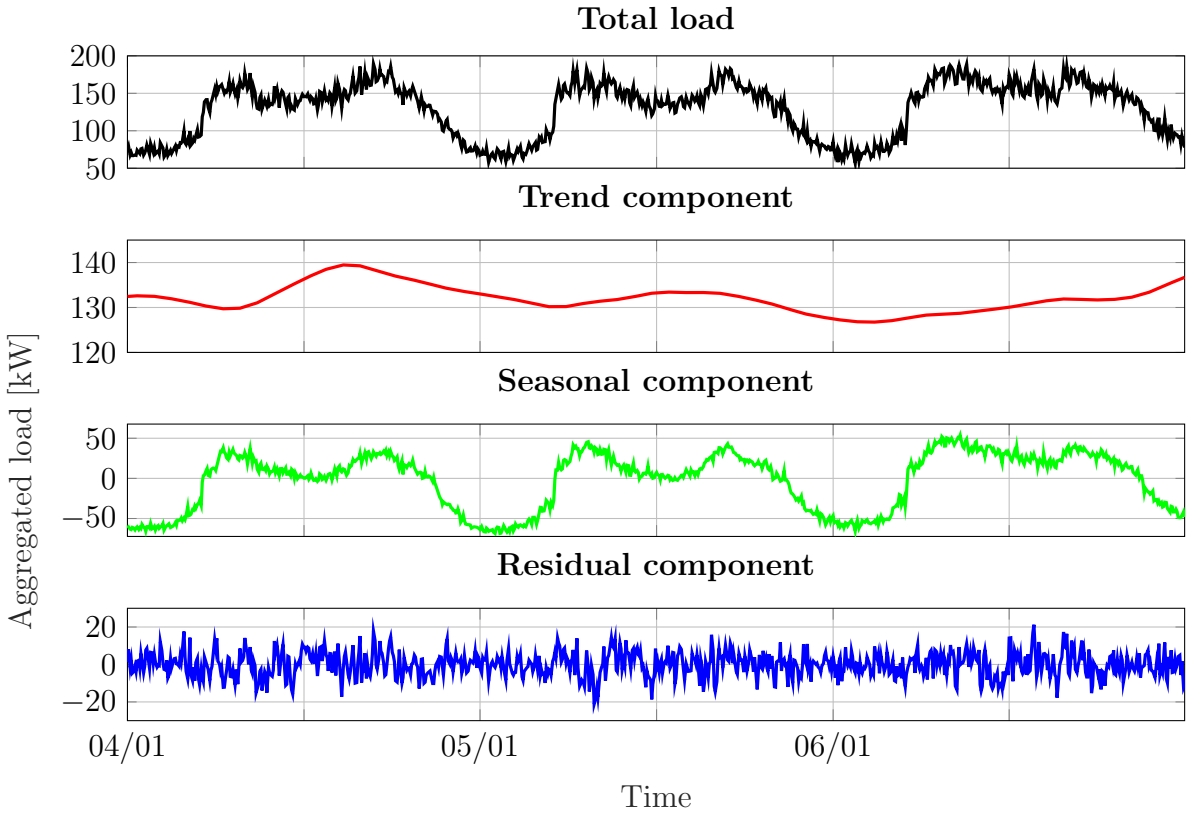

Figure 6: Example of decomposition for 138 households and a period of three days in January 2018.

An example of load decomposition is shown in Fig. 6. To calculate the trend component at time t^*, polynomials are fitted locally to all data points n_{tr} time steps before and after t^*, using weights which are inversely proportional to the distance from t^*. Hence, n_{tr} is an input parameter for the decomposition method, which defines how smooth the resulting trend component is, with larger values of n_{tr} resulting in smoother trends. The original time series $\boldsymbol{w}$ can then be detrended by subtracting $\boldsymbol{T}$

$$\hat{w}_t = w_t - T_t, \quad \forall\ t \in \{1, ..., N_{\text{tot}}\}. \tag{8}$$

Next, $\hat{w}_t$ is used to calculate the seasonal component. The duration of the seasonal electricity consumption pattern is equal to one week, corresponding to $\gamma = 2016$ time steps for a 5−minute resolution. If ζ is the number of weeks contained in $\hat{\boldsymbol{w}}$, then $\hat{\boldsymbol{w}}$ is divided into γ sub-time series $\boldsymbol{S}^{*,i}$. Each sub-time series $\boldsymbol{S}^{*,i}$, of length ζ, consists of the power values of the same time steps of each week, such that $S^{*,i} = \{\hat{w}_i, \hat{w}_{i+\gamma}, \ldots, \hat{w}_{i+(\zeta-1)\gamma}\}$, with $i \in \{1, 2, ..., \gamma\}$. Again polynomials are fitted locally, however, now fitted only to the sub-time series, with the parameter n_{s} defining how many time steps before and after are taken into account. After the seasonal component is calculated, the residual component $\boldsymbol{R}$ is defined as

$$R_t = w_t - T_t - S_t, \quad \forall\ t \in \{1, ..., N_{\text{tot}}\}. \tag{9}$$

.

3.3.2. Linear regression on residual consumption

This baseline method relies on both load decomposition and BR. First, the clean time series $\boldsymbol{w}$ and the residual consumption are obtained, as previously described. The residual consumption is correlated with the time of day, external temperature and solar radiation. For this reason a multiple LR is used, with the residual being the dependent variable, and the other three factors being the independent variables $\boldsymbol{X}$, also referred to as regressors. Coefficients $\boldsymbol{\beta}$ link the independent variables with the dependent variable, whereas $\boldsymbol{\epsilon}$ represents the error between the regression and the real values. The relationship between $\boldsymbol{R}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ for b data points is described in a matrix form by:

$$\boldsymbol{R} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{10}$$

$$\boldsymbol{R} = \begin{bmatrix} R_1 \\ R_2 \\ \vdots \\ R_\mathrm{b} \end{bmatrix}, \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,\mathrm{p}} \\ 1 & x_{2,1} & \cdots & x_{2,\mathrm{p}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{\mathrm{b},1} & \cdots & x_{\mathrm{b},\mathrm{p}} \end{bmatrix}, \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_\mathrm{p} \end{bmatrix}, \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_\mathrm{b} \end{bmatrix}. \tag{11}$$

A total number of p predictors are used: ambient temperature (in Celsius), solar radiation (kW/m^2), and v pairs of time predictors. Column $x_{:,1}$ contains the b ambient temperature values, whereas column $x_{:,2}$ contains the solar radiation values. Columns $x_{:,3}$ to $x_{:,\mathrm{p}}$ contain the sine/cosine transformed time values, as explained below.

Meter data is recorded every 5 minutes, therefore the time stamp g_i associated with data point i is mapped with the help of a strictly increasing function $\mathbf{f} : \mathbb{R} \to \mathbb{R}$ to a range $[0, 24)$, representing the hour of day. However, the hour of the day should not be directly included as an input to the LR, since its influence on residential power consumption is not linear. The time stamp of each data point is therefore transformed for each pair m out of v time predictor pairs as:

$$x_{i,2m+1}\left(f\left(g_i\right)\right) = \sin\left(m\frac{2\pi}{24}f\left(g_i\right)\right), \ \forall m \in \{0, ..., \mathrm{v}\}, \ \forall i \in \{1, ..., \mathrm{b}\}. \tag{12}$$

$$x_{i,2m+2}\left(f\left(g_i\right)\right) = \cos\left(m\frac{2\pi}{24}f\left(g_i\right)\right), \ \forall m \in \{0, ..., \mathrm{v}\}, \ \forall i \in \{1, ..., \mathrm{b}\}. \tag{13}$$

Note, that $x_{i,j}(24) = x_{i,j}(0)$. Finally, the least squares method is applied on (10) to obtain coefficients $\boldsymbol{\beta}$ which minimize the squared sum of errors $\boldsymbol{\epsilon}$.

3.3.3. Hybrid method

A hybrid method relying on the linear combination of the previously described baseline methods is also used. It is based on the assumption that each of the methods contains useful information, which can be combined to form a more accurate baseline. Let $\boldsymbol{u}^j$ denote the

estimation of baseline method j, as listed in Table 1. The estimation of the hybrid method over a period N_{d} is calculated as a linear combination of all individual $\boldsymbol{u}^j$ estimations as

$$u_i^{\mathrm{hybrid}} = \sum_{j=1}^{5} \alpha_{j,t} u_i^j, \ \forall \ i \in \{1, ..., N_{\mathrm{d}}\}. \tag{14}$$

To calculate coefficients $\alpha_{j,i}$, first a training set of evaluation periods is chosen randomly, so that they do not coincide with the DR experiments. Next, the different methods are used to produce baseline estimations for each evaluation period. These estimates are used in (14), to produce the hybrid method estimate. Finally, a least-squares fit on the residuals against the real consumption is applied for each i separately.

4. Baseline Methods Evaluation

In this section the results of the evaluation of the different baseline methods are presented. Similarly to the data replacement method, the performance of the different baseline methods was evaluated on time periods where no DR experiments were conducted. The duration of each test is again 36 time steps, which translates to a 3-hour test period. First, the clean aggregated time series $\boldsymbol{w}$ is obtained, as described in subsection 3.2. Next, the data set is split into a training/cross validation set (80% of the data set) and a test set (20% of the data set). All methods, except for LI where it is unnecessary, were trained and optimized on the training set using 10-fold cross validation. Finally, all methods were tested on the test set. The decomposition-based models M3, M4 and M5 were found to perform best with the parameters $(n_{\mathrm{s}}, n_{\mathrm{tr}}) = (7, 55)$, $(n_{\mathrm{s}}, n_{\mathrm{tr}}) = (7, 87)$ and $(n_{\mathrm{s}}, n_{\mathrm{tr}}) = (7, 137)$, respectively. The results of the six methods are summarized in Table 2, with the tests being carried out using the aggregated load of the full portfolio of 138 households.

Method	STD [kW]	IQR_{95} [kW]	IQR_{99} [kW]
M1 - LI	14.21	57.94	76.29
M2 - FBA	12.23	51.38	67.57
M3 - D - R	11.68	47.22	65.01
M4 - D + FBA on R	11.48	46.78	63.36
M5 - D + LR on R	11.43	46.40	63.22
M6 - Hybrid	11.01	44.52	61.27
MA Benchmark	6.48	26.48	37.28

Table 2: Overview of the different baseline method results for 138 households.

LI is used as the upper benchmark for comparisons, whereas MA is used as a lower benchmark. FBA outperforms LI in all metrics, with a 13% reduction of the errors STD, a 12% reduction of IQR_{95} and a 11% reduction of IQR_{99}. All three variants of load decomposition outperform the simple FBA, showing that the use of load decomposition can result in smaller baseline estimation errors. LR on the residual consumption gives the best result,

whereas the hybrid method achieves a further reduction of the errors. More specifically, compared to LI, it achieves a 23% reduction of the errors STD, a 23% reduction of IQR_{95} and a 20% reduction of IQR_{99}. In Fig. 7 the probability distribution of the error of the hybrid method for 300 test periods is shown.

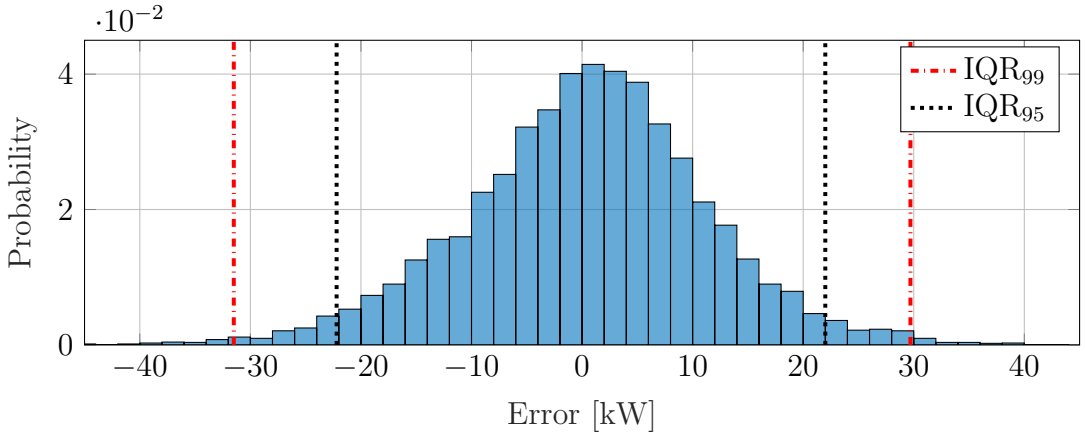

Figure 7: Relative frequency histogram of the error distribution of the hybrid method for 300 test periods.

The performance of the methods is not the same for each time step. As expected, all methods perform better at the first and last time steps of the estimation. This is depicted in Fig. 8, where the STD of the error for each of the 36 time steps for all different baseline methods is shown. The graph also reveals the much larger variability of the LI's performance, which is concealed by the average error results of Table 2. The hybrid method performs better than all the other five methods for all 36 time steps, and with a smaller variability of its performance.

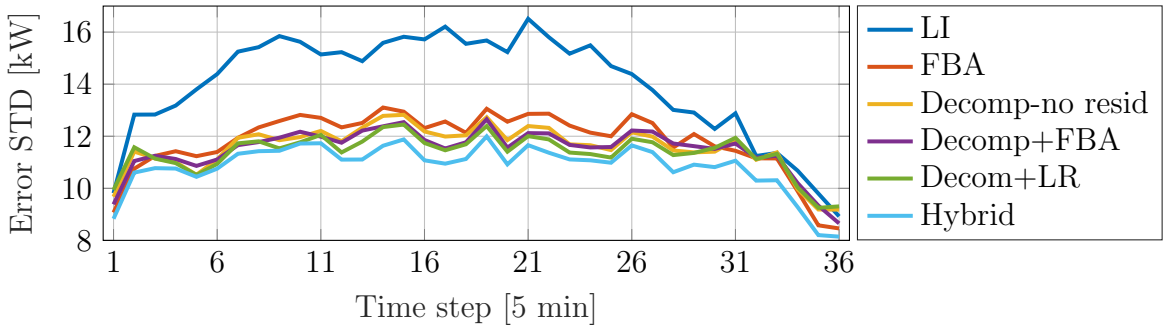

Figure 8: STD of the error term for each of the 36 time steps and for all different baseline methods.

Finally, the effect of the aggregation size on the STD of the errors is analyzed. To this end, the portfolio size was reduced by randomly choosing a subset of the population, and the calculations were repeated for 116, 94, 72 and 50 households. In general, the results

depend on the composition of the chosen subset. To isolate the effect of the aggregation size, the subsets were repeatedly (and randomly) chosen, until the mean of the calculated errors converged to a steady-state value. The results are shown in Fig 9.

The total error increases with the aggregation size. This is expected, because of the nature of the error metric. When more households are considered, the aggregated load has larger values, which results in larger errors on average. However, the error per household decreases as the aggregation size increases. This reflects the impact of the aggregation size on the natural smoothing of the aggregated load, and the reduction of the impact of customer uncertainty per household. As we will show in the following section, an error STD of approximately 0.1 kW per household is significantly smaller than the contribution of each household in the load reduction experiments. This allows us to quantify flexibility with acceptable accuracy.

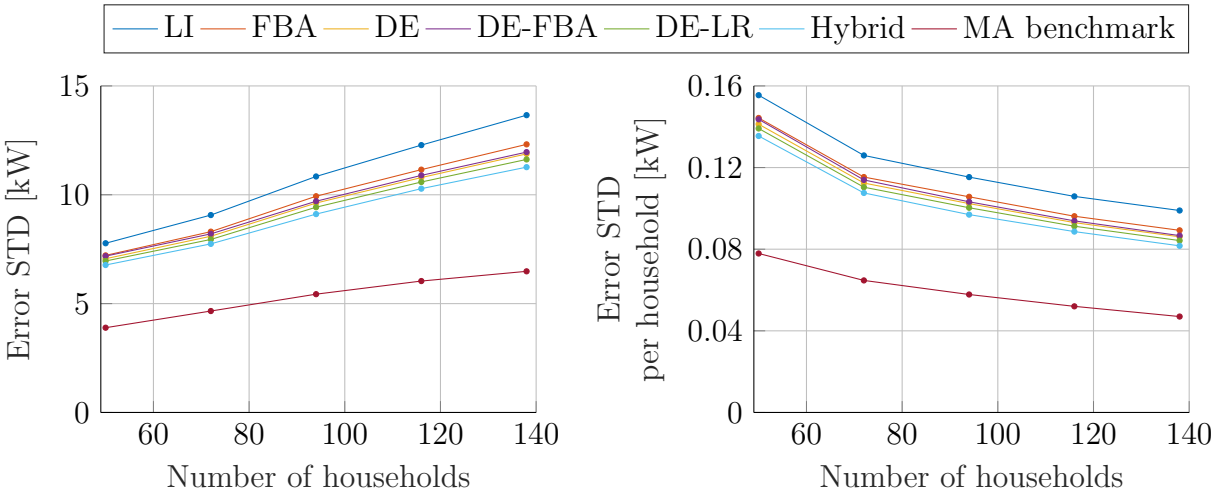

Figure 9: STD of total error (left) and error per household (right) as a function of the aggregation size.

5. Flexibility Identification

In order to offer flexibility services, aggregators need to estimate the available flexibility in their portfolio. Accordingly, a number of flexibility identification experiments were carried out in the context of EcoGrid 2.0. A total of 90 DR experiments were carried out, with aggregation sizes varying between 50 and 138 households. This section describes how these experiments were used to create a flexibility model, which models the portfolio's response depending on time of day and ambient temperature. The flexibility model contains two parts. The response model, describing the loads response during the load reduction phase, is presented in subsection 5.1. The rebound model, describing the loads rebound, is presented in subsection 5.2.

5.1. Response model

First, the process of the experiments evaluation, and the proposed response flexibility model, are described in subsection 5.1.1. Second, the validation of this model is presented in 5.1.2. Finally, the impact of the aggregation size and the random selection of households is assessed in 5.1.3.

5.1.1. Evaluation of experiments and flexibility response model

A typical result of a flexibility experiment is depicted in Fig. 10. The upper graph shows the actual measured consumption in blue color, whereas the red curve represents the estimated baseline using the hybrid method. The lower graph depicts the load deviation, i.e., the difference between the baseline and the actual load, together with the IQR_{95} baseline uncertainty. The one-hour load reduction period is highlighted by the light orange background, and the evaluation extends to a subsequent two-hour rebound period.

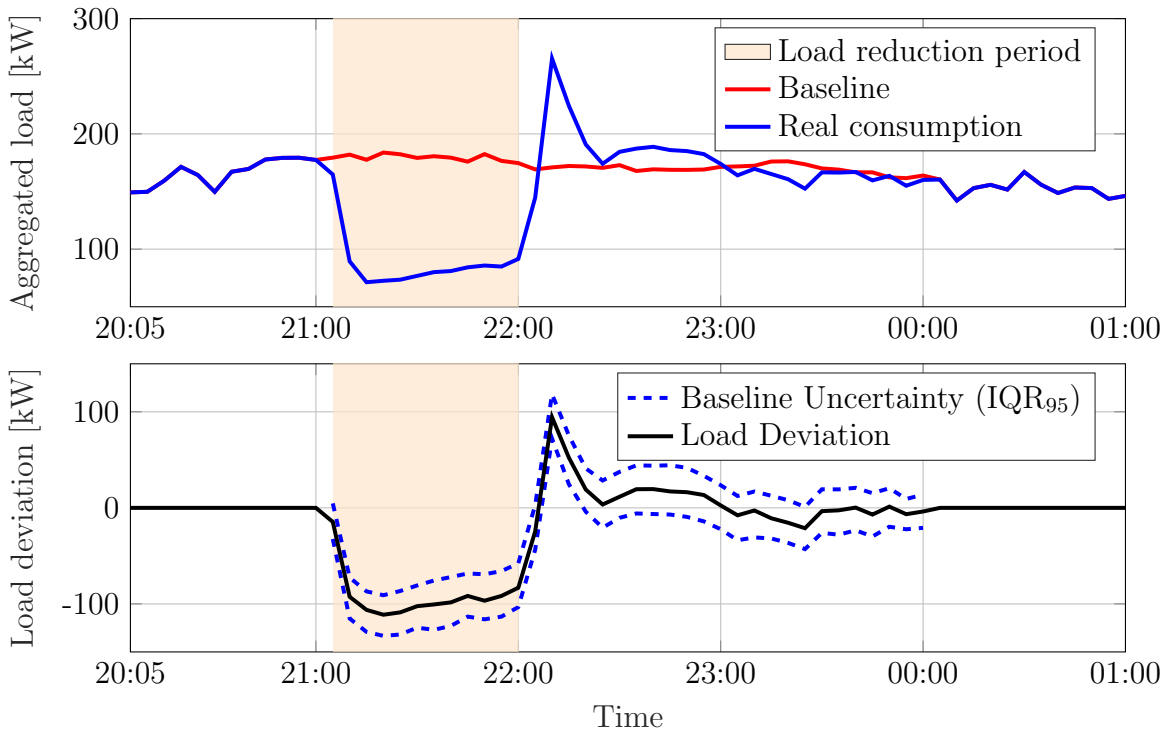

Figure 10: Experiment result example - Top: Baseline (red) and measured consumption (blue), Bottom: Load deviation and the IQR_{95} of the baseline uncertainty.

Each of the conducted experiments involved a different number of households, and the load reduction period was not always the same. The average load reduction per household can be used to quantify flexibility, given the relatively constant load reduction observed in the experiments (see Fig. 10). This allows us to calculate a flexibility indicator in a normalized manner, irrespective of the number of households or the experiment's duration. For the evaluation of each experiment, three steps are followed:

1. *Obtain aggregated load.* The aggregated load $\boldsymbol{y}$ of the n_l households participating in the $l-$th experiment is obtained.
2. *Create baseline.* A baseline is estimated as described in section 3, for the corresponding n_l households, and for a duration N_l equal to three times the load reduction period d_l. In other words, $N_l = 3\ d_l$.
3. *Calculate average response.* Using the estimated baseline, the average response per household is calculated. In the example of Fig. 10, the load reduction occurs between 21.00 and 22.00, with 21.05 being the first time step of the evaluation. Due to the relatively slow ICT infrastructure, loads respond with a delay. To avoid underestimating flexibility, the first time step is omitted from the calculation of the average response.

For each experiment l with a starting time t_l, n_l participating loads, and a load reduction period equal to d_l, the average response P_l^{res} is calculated as

$$P_l^{\text{res}} = \frac{\sum_{i=2}^{i=d_l} u_{t_l+i} - y_{t_l+i}}{n_l\ (d_l - 1)}. \tag{15}$$

As pointed out in [30], the baseline accuracy can have an effect on the calculation of flexibility. Each P_l^{res} value is calculated by averaging the estimated load deviation over $d_l - 1$ time steps. To assess the effect of possible baseline errors on the average response, a new offset error metric e^{off} is introduced. A positive offset error e^{off} will result in overestimation of flexibility, whereas a negative error in underestimation. To quantify the performance of the baseline method in terms of offset errors, an evaluation procedure similar to the one followed for the instantaneous error e was carried out. 300 evaluation periods with a duration of three hours ($N_k = 36$) were chosen randomly, and the evaluation process was repeated for different aggregation sizes. The offset error per household for each evaluation period is calculated as

$$e_k^{\text{off}} = \frac{\sum_{i=2}^{i=d_k} u_{t_k+i}^k - y_{t_k+i}}{n_k\ (d_k - 1)}. \tag{16}$$

Note that the offset error is normalized per household, and that it is calculated for time steps $2, \ldots, d_k$, which are also the time steps used for the calculation of P^{res}. Identical results were obtained by evaluating e^{off} for $N_k = 36$ or $N_k = 18$. However, e^{off} was found to exhibit a strong dependence on the number of houses, as is the case for the instantaneous error e. In Fig. 11, the STD of e^{off} is shown. Notice the very small values for the benchmark errors; this is explained by the MA nature of the benchmark estimation, that results in negligible average errors, which is not the case for instantaneous errors.

After calculating P_l^{res} for each experiment, these values are fitted with a flexibility model. The expected response $\hat{P}^{\text{res}}$ per household is expressed as a function of the hour of day h and ambient temperature T^{amb} as

$$\hat{P}^{\text{res}}(h, T^{\text{amb}}) = \lambda_1 \cos\left(\frac{2\pi h}{24}\right) + \lambda_2 \sin\left(\frac{2\pi h}{24}\right) + \lambda_3 T^{\text{amb}} + \lambda_4. \tag{17}$$

The hour of day is projected onto the interval $[-1, 1]$ using the sine/cosine functions. This model design decision was made to retain the continuity of flexibility, since the model's

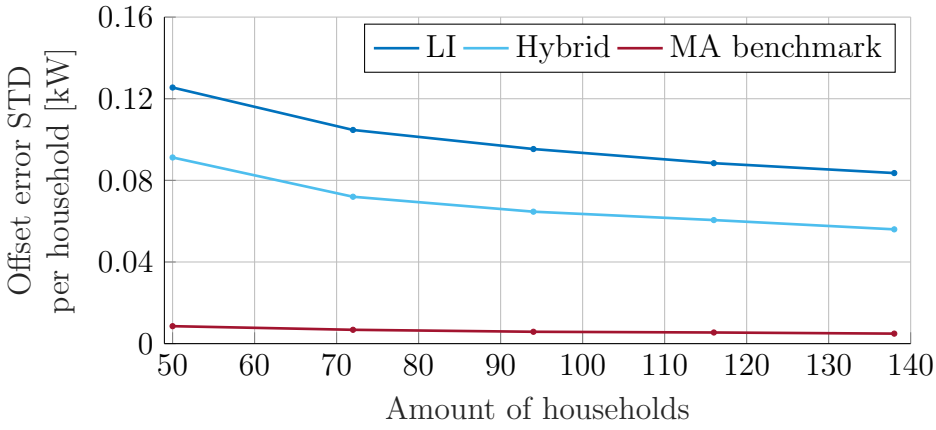

Figure 11: STD of the offset error per household, as a function of the aggregation size.

predictions for $h = 0$ have to be equal to $h = 24$. Moreover, experimental data indicated a linear dependency between the load reduction and the ambient temperature. The model parameters λ_1, λ_2, λ_3 and λ_4 are calculated using a weighted least square method. As shown in Fig. 11, the evaluation of experiments involving fewer houses leads, in general, to larger offset errors per household. As a result, the estimated P_l^{res} value will be less reliable. If the offset error STD corresponding to an experiment l is denoted by e_l^{off}, then for each experiment the following weight ω_l is assigned

$$\omega_l = \frac{1/e_l^{\text{off}}}{\sum_l (1/e_l^{\text{off}})}. \tag{18}$$

5.1.2. Model validation

80 of the total 90 experiments were randomly chosen as training data to fit the flexibility model, and the remaining 10 experiments were used for validation. The evaluation results upon the training data were used to calculate the coefficients of model (17), using weighted least squares fitting. The results of this fitting are summarized in Table 3.

Parameter	Value	Unit
λ_1	0.125	$[kW]$
λ_2	-0.007	$[kW]$
λ_3	0.039	$[kW/^oC]$
λ_4	-0.739	$[kW]$

Table 3: Fitted values of the coefficients of model (17).

The fitted model is shown in Fig. 12, where each point in the graph represents the average load reduction per household during an experiment, and the additional color dimension indicates the average ambient temperature during the experiment. The fitted model is depicted through isotherm lines, for temperatures of -10°C, -5°C, 0°C, 5°C and 10°C.

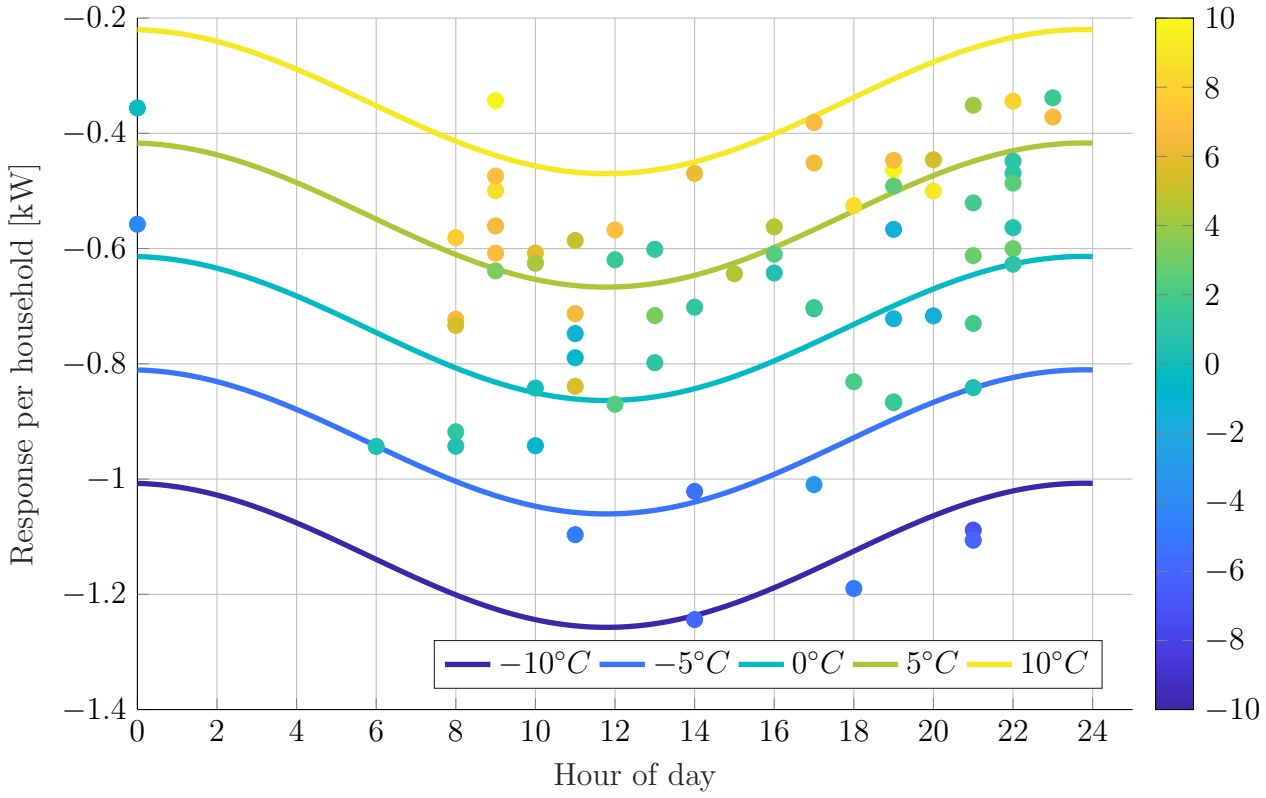

Figure 12: Fitted flexibility model as a function of time of day and ambient temperature. Each point represents the average load reduction per household during an experiment.

One can notice that even for the same hour of day and similar temperatures, the resulting response varies considerably. This is attributed to the baseline uncertainty, the varying portfolio composition throughout the experiments, and the various customer-induced disturbances, among other factors. The residuals of the fitted model are approximately normally distributed, with a standard deviation of 0.12 kW. Using a more complicated fitting model may yield smaller residuals in the training data, but the resulting model will not be necessarily better. The reason for this is that the residuals also reflect the uncertainty of flexibility provision, observed throughout the experiments. Trying to minimize the residuals often leads to an over-fitted model, which produces less accurate flexibility estimations - in other words, worse validation results. This is the reason why a simpler function for the flexibility model was chosen, because it produced better results on the validation data set. The results of the validation are summarized in Table 4. Most of the responses estimated via (17) are close to the measured values. In only two cases a mismatch greater than 0.1 kW per household was observed. Given the STD of the residuals and the uncertainty of the offset error (see Fig. 11), such mismatches are expected.

number of experiment	predicted response	measured response	estimation mismatch	amount of households
1	-1.02	-1.10	0.074	75
2	-0.60	-0.56	-0.036	85
3	-0.36	-0.37	0.008	74
4	-0.81	-0.80	-0.012	75
5	-0.76	-0.94	0.179	109
6	-0.55	-0.50	-0.048	112
7	-0.57	-0.52	-0.054	116
8	-0.77	-0.70	-0.070	114
9	-0.51	-0.40	-0.112	73
10	-0.68	-0.62	-0.065	76

Table 4: Results of the validation of model (17) for 10 experiments.

5.1.3. Uncertainty assessment of aggregation size

As explained in the previous subsection, the amount of load reduction cannot be fully described by the hour of day and the ambient temperature. Other factors, such as user-induced disturbances or weather data inaccuracies, are introduced in the model as uncertainties. In this subsection we focus on the effect of the activation of a subset of the whole portfolio, on the delivered load reduction. The flexibility model expressed via (17) estimates the load reduction per household for the whole available portfolio. However, activating a randomly selected subaggregation may result in a different response per household, compared to the one obtained from all loads. To assess the effect of random selection, the marginal contribution of each load to the load reduction was calculated.

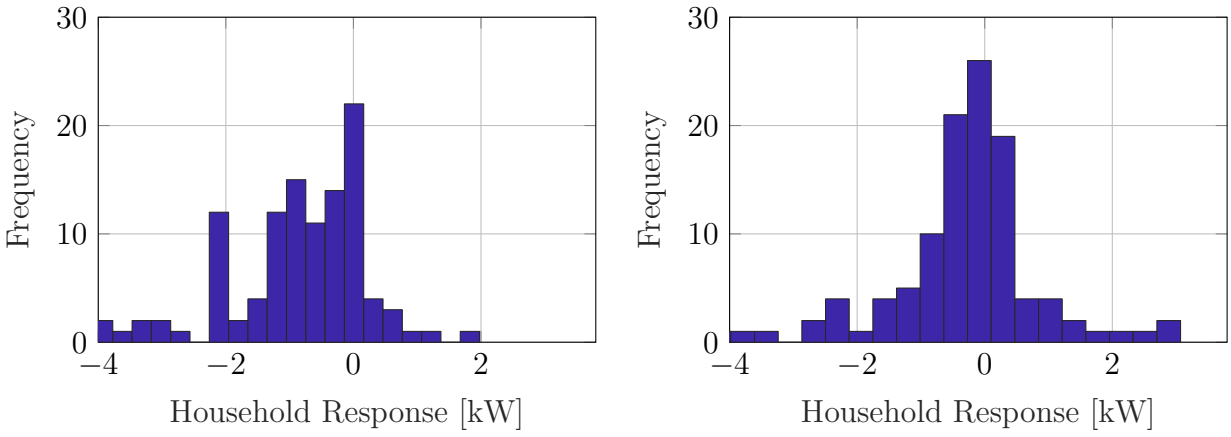

Figure 13: Frequency histograms of the estimated marginal contribution of a single household - (results for experiment 1 are shown on the left subplot, and for experiment 2 on the right subplot).

Two load reduction experiments with the same 110 participating loads were chosen for

the analysis. The following process was followed for both experiments. First, the total load reduction during the experiment was calculated. Afterwards, one of the 110 loads was removed, and the load reduction was recalculated for the remaining 109 loads. The difference of these two values gives the contribution of a particular household to the overall load reduction. This process was repeated for each load, resulting in 110 contribution values. The results of the marginal contributions for these two experiments are shown in Fig. 13.

Two interesting observations can be made by examining the histograms of the marginal contributions. First, some of the loads appear to have positive contributions to the load reduction. This is mainly attributed to the uncertainty of the baseline. If a relatively large increase of the uncontrollable consumption of a specific household during the experiment is not captured by the baseline estimation, then its marginal contribution will appear as positive. Similarly, a decrease of the uncontrollable consumption not captured by the baseline estimation may inflate the actual contribution. Second, there is a wide distribution of the marginal contributions among the population.

To assess the impact of the aggregation size, we randomly constructed subaggregations of varying sizes. For each subaggregation, the marginal contributions of the loads were averaged. This process was repeated 150 times for each subaggregation size, and the results are presented in Fig. 14 and Fig. 15 for experiment 1 and 2 respectively.

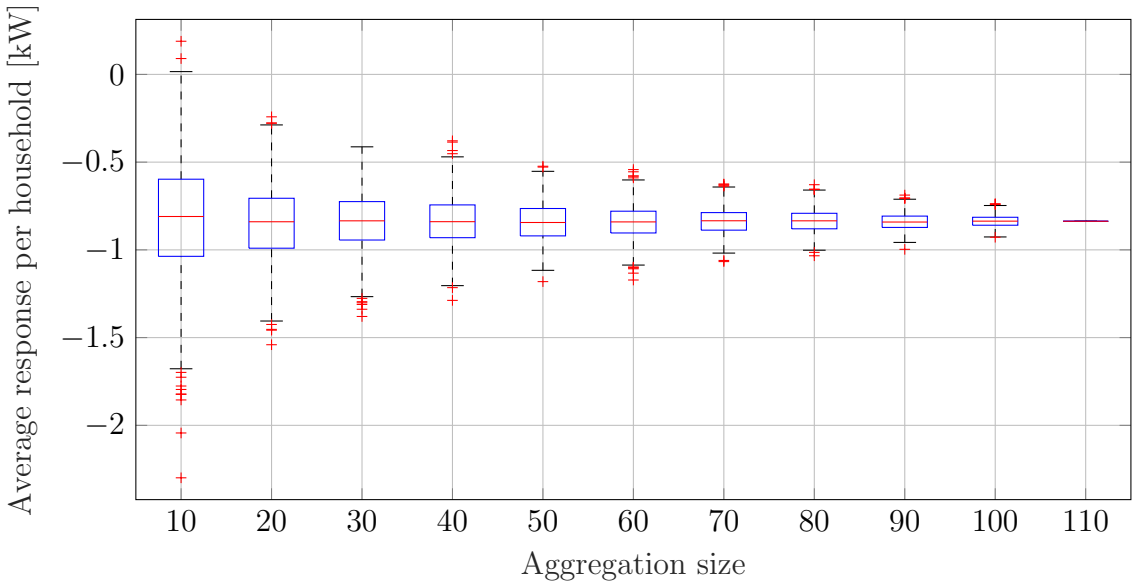

Figure 14: Boxplot of the average response per household for varying aggregation sizes: experiment 1.

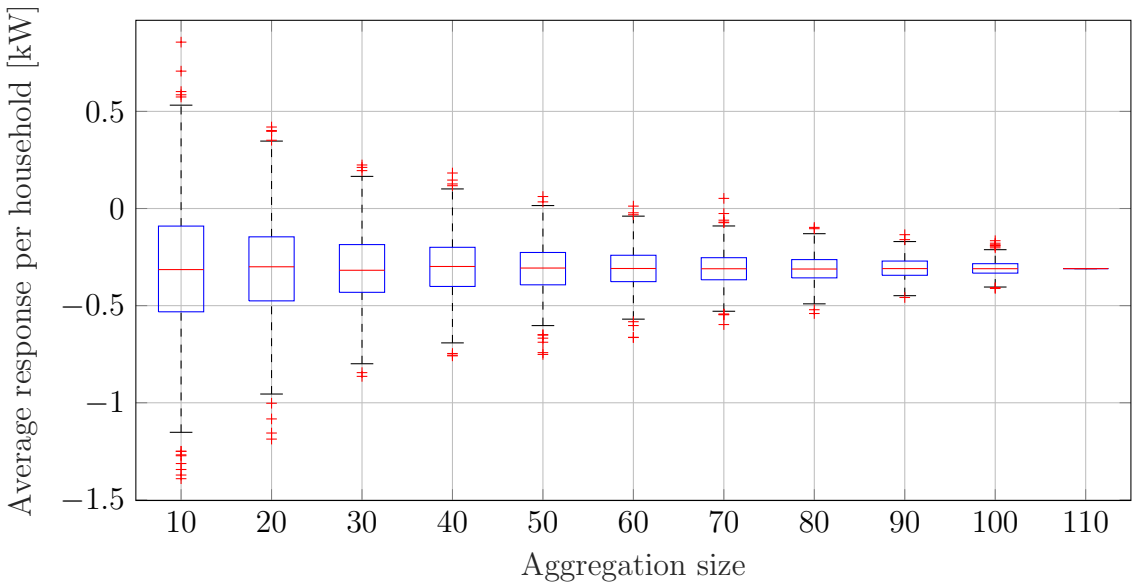

Figure 15: Boxplot of the average response per household for varying aggregation sizes: experiment 2.

The expected response per household remains constant, regardless of the subaggregation size. However, the results for both experiments indicate a logarithmic decrease of the uncertainty in the response, as the number of loads increases. In the first experiment, the estimated response for the 110 households is equal to 0.84 kW per household. Fig. 14 reveals that if the same experiment was conducted with 70 loads, then the response would lie in the range of $[0.79 - 0.89]$ kW, with a probability of 50 %. The response per household could be as low as 0.63 kW, or as high as 1.05 kW per household. Notice that these ranges are similar to the uncertainty ranges introduced by the baseline offset error in Fig. 11, which is why P_l^{res} values were weighted by ω_l, according to the number of participating houses.

Similar uncertainty ranges were found for the second experiment, as seen in Fig. 15. Even if the estimated response for the 110 households is equal to 0.31 kW per household, the ranges of the responses obtained by smaller aggregations exhibit similar variability. These uncertainty ranges can be used by the aggregator to deliver a service by activating the necessary amount of households, considering the baseline uncertainty and the service specifications.

5.2. Rebound

A large number of experiments resulted in a similar rebound behavior, i.e., relatively short in duration but with a large peak value. In this subsection, a model describing the rebound behavior of the loads will be presented. For this purpose, only tests with a duration equal to one hour are considered. The normalized load deviation during an experiment $P_{l,i}^{\text{nor}}$ is introduced, which is defined as the load deviation per household, normalized against the

average load reduction P_l^{res}. $P_{l,i}^{\text{nor}}$ is calculated as

$$P_{l,i}^{\text{nor}} = \frac{u_{t_l+i} - y_{t_l+i}}{n_l P_l^{\text{res}}}, \quad \forall\, i \in \{1, ..., N_{\text{d}}\}. \tag{19}$$

To obtain a representative rebound behavior, the normalized responses of $n_{\text{b}} = 25$ experiments were averaged, when all loads were released immediately after the load reduction period. The contribution of each experiment is weighted by a weight ρ_l. This weight is similar to ω_l, but is calculated considering the instantaneous error distributions of e (see Fig. 9), because $P_{l,i}^{\text{nor}}$ is calculated for each time step i. If the instantaneous error STD corresponding to an experiment l is denoted by e_l^{inst}, then for each experiment the following weight ρ_l is assigned

$$\rho_l = \frac{1/e_l^{\text{inst}}}{\sum_l (1/e_l^{\text{inst}})}. \tag{20}$$

The average normalized load deviation $\overline{P}_i^{\text{nor}}$ is calculated as

$$\overline{P}_i^{\text{nor}} = \sum_{l=1}^{n_{\text{b}}} \rho_l P_{l,i}^{\text{nor}}, \quad \forall\, i \in \{1, ..., N_{\text{d}}\}. \tag{21}$$

The expected normalized load behavior when all loads are released after the load reduction period is shown in Fig. 16. A simple rebound-shaping strategy is described next, which can significantly reduce the large rebound peak power after the end of the load reduction period.

We introduce a_i as the share of released loads at step i, with $a_i \in [0, 1]$, and $\sum_{i=1}^{N_{\text{d}}} a_i = 1$. If a share of the loads a_{13}, instead of the whole portfolio, is released at step $i = 13$, then these loads will exhibit a rebound behaviour as shown with blue color in Fig. 16. Note that by convention a release at $i = 13$ refers to release commands sent at the end of the load reduction period. We assume that the contribution of the rest of the loads $(1 - a_{13})$ in the load reduction will be the same. The shaped load deviation P_i^{sh}, achieved by applying a gradual load release, can be calculated as

$$P_i^{\text{sh}} = \begin{cases} \overline{P}_i^{\text{nor}}, & \forall\, i \in \{1, \ldots, 12\} \\ \overline{P}_{12}^{\text{nor}} \left(1 - \sum_{r=13}^{i} a_r\right) + \sum_{r=0}^{i-13} a_{i-r} P_{13+r}^{\text{sh}}, & \forall\, i \in \{13, \ldots, 36\}. \end{cases} \tag{22}$$

The accuracy of model $\overline{P}_i^{\text{nor}}$ to describe the rebound behavior was validated by performing a series of gradual release experiments. After one hour of load reduction, random subsets of the portfolio were released according to five different release strategies, as summarized in Table 5. The percentage column shows the shares of release, i.e., $[50, 25, 25]$ means that first 50% of the loads were released, followed by a 25% share and another 25% in the end. Column time shows the time in minutes, when the commands for gradual release were sent by the aggregator to the loads. The sequence has the end of the load reduction period as time reference, which was 1 hour after the experiment commenced. Therefore, a sequence of $[0, 10, 20]$ means that the first share of loads were released right after the end of the load

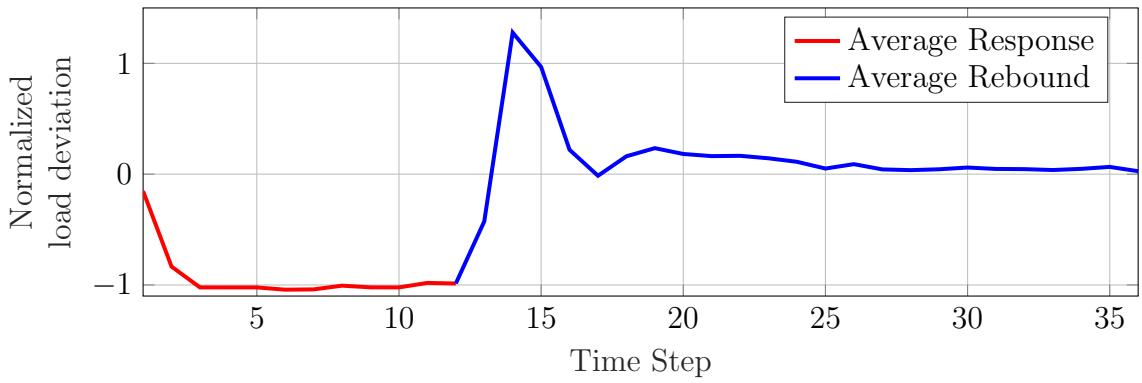

Figure 16: Average normalized load deviation $\overline{P}_i^{\text{nor}}$ from 25 load reduction experiments.

reduction period. The second share was released 10 minutes later, and another 10 minutes later the last loads were released. With reference to the shares a_i of (22), this results in $a_{13} = 0.5$, $a_{15} = 0.25$ and $a_{17} = 0.25$.

Test	Percentage	Time [min]
(a)	[50,10,10,10,10,10]	[0,10,20,30,40,50]
(b)	[50,25,25]	[0,15,30]
(c)	[50,25,25]	[0,15,30]
(d)	[50,25,25]	[0,10,20]
(e)	[20,20,20,20,20]	[0,10,20,30,40]

Table 5: Description of gradual release experiments.

In Fig. 17 the results from these five experiments are shown. The actual response is shown with red color, whereas the calculated response via (22) is shown with blue color. In all cases the actual response of the loads under gradual releases was found to be close to the response calculated by using (22). Such a strategy can be used to mitigate the high rebound peak, and shape the rebound profile according to the aggregator's objectives (for instance to provide a service to the DSO).

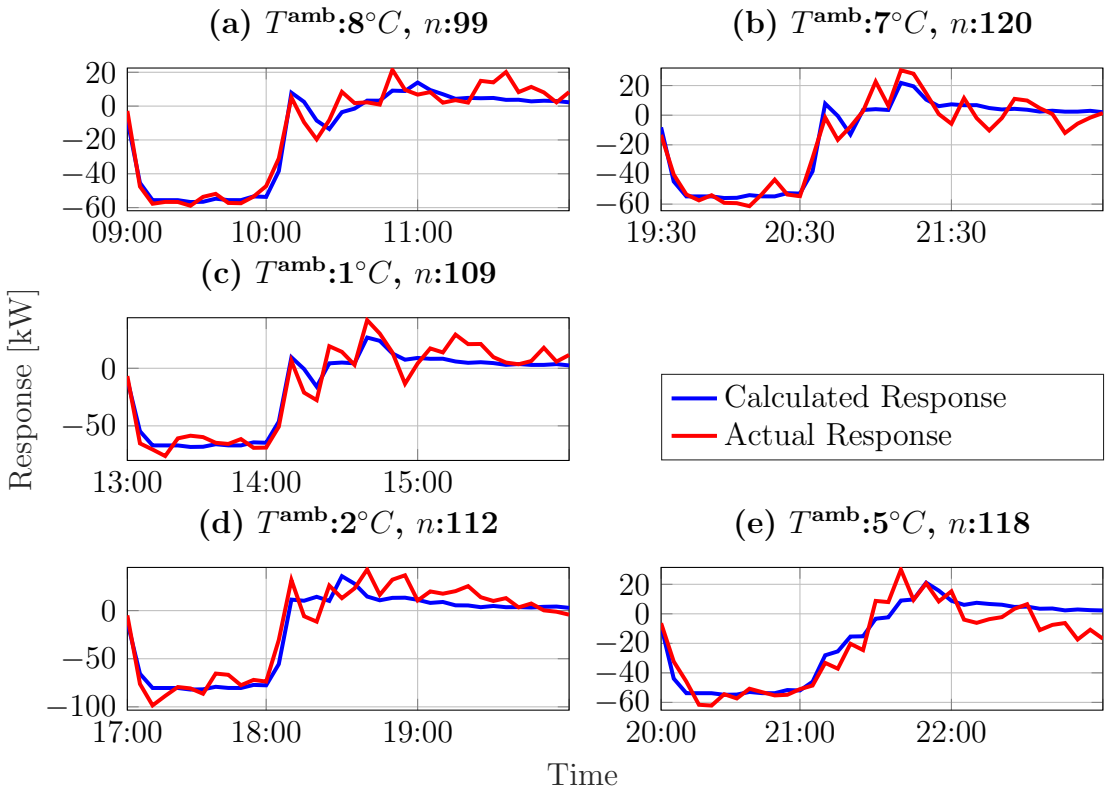

Figure 17: Calculated and actual response for five gradual release tests with varying portfolio size and ambient temperature.

6. Conclusion

In this paper we proposed a method for identifying the aggregated flexibility of residential thermal loads by evaluating DR activations, based only aggregated behind the meter measurements and weather data. For this purpose, six different baseline methods were presented and compared. Results show that load decomposition and auto-regressive models lead to significant improvements with relation to linear interpolation. A hybrid method, combining the benefits of regression and decomposition, was found to further increase the accuracy of the baselines.

Results showed that the standard deviation of the baseline offset error ranges from 0.055 to 0.075 kW per household, for aggregation sizes ranging from 138 of 70 households, respectively. The amount of achievable load reduction was found to vary between 0 kW and -1.2 kW per household, for a large number of experiments involving $75 - 115$ households. The considerably smaller baseline errors, compared to the calculated load reductions, allows the identification of flexibility with good accuracy. The available load reduction per household was modelled as a function of the ambient temperature and time of day, using the DR evaluations as training data. The standard deviation of the residuals of the model was found to be 0.12 kW per household. This uncertainty is higher than the baseline uncertainty, given

that it also includes the uncertainty of random customer behavior, as well as the effect of random selection of households participating in a DR experiment. We assessed the effect of random selection of households by evaluating their individual flexibility contributions during two DR experiments involving 110 households. The response of a subaggregation remained on expectation constant, as the subaggregation size decreased. However, the uncertainty in the response scaled faster than than the baseline uncertainty, resulting in very large response ranges for subaggregation below 50 households.

The rebound behavior of the loads was also modelled, using a set of 25 experiments. This model can be used to predict the rebound behavior of the aggregation under gradual load releases, and mitigate the large rebound peaks. The accuracy of both the response and the rebound models were validated with real experiments. Future work will investigate if clustering the households can result in a reduction of the uncertainty when activating small subaggregations.

Acknowledgement

The authors would like to acknowledge the financial support of the EUDP project, EcoGrid 2.0, No 64015-002.

Reference

[1] D. Fraile and A. Mbistrova, "Wind in power 2017 - Annual combined onshore and offshore wind energy statistics," tech. rep., Wind Europe, 2018.

[2] Energinet, "Environmental Report 2017," tech. rep., Energinet, Fredericia, 2017.

[3] Q. Wang, C. Zhang, Y. Ding, G. Xydis, J. Wang, and J. Østergaard, "Review of real-time electricity markets for integrating distributed energy resources and demand response," *Applied Energy*, vol. 138, pp. 695–706, 2015.

[4] K. Spiliotis, A. I. R. Gutierrez, and R. Belmans, "Demand flexibility versus physical network expansions in distribution grids," *Applied Energy*, vol. 182, pp. 613 – 624, 2016.

[5] C. Zhang, D. Yi, N. C. Nordentoft, P. Pinson, and J. Østergaard, "Flech: A danish market solution for dso congestion management through der flexibility services," *Journal of Modern Power Systems and Clean Energy*, vol. 2, no. 2, pp. 126–133, 2014.

[6] C. Zhang, Q. Wang, J. Wang, P. Pinson, J. M. Morales, and J. Østergaard, "Real-time procurement strategies of a proactive distribution company with aggregator-based demand response," *IEEE Transactions on Smart Grid*, vol. 9, no. 2, pp. 766–776, 2018.

[7] P. Olivella-Rosell, E. Bullich-Massagué, M. Aragüés-Peñalba, A. Sumper, S. Ø. Ottesen, J.-A. Vidal-Clos, and R. Villafáfila-Robles, "Optimization problem for meeting distribution system operator requests in local flexibility markets with distributed energy resources," *Applied Energy*, vol. 210, pp. 881–895, 2018.

[8] C. Rosen and R. Madlener, "Regulatory options for local reserve energy markets: Implications for prosumers, utilities, and other stakeholders," *The Energy Journal*, 2014.

[9] A. Roos, "Designing a joint market for procurement of transmission and distribution system services from demand flexibility," *Renewable Energy Focus*, vol. 21, pp. 16–24, 2017.

[10] G. Bianchini, M. Casini, A. Vicino, and D. Zarrilli, "Demand-response in building heating systems: A Model Predictive Control approach," *Applied Energy*, vol. 168, pp. 159–170, 2016.

[11] J. Le Dréau and P. Heiselberg, "Energy flexibility of residential buildings using short term heat storage in the thermal mass," *Energy*, vol. 111, pp. 991–1002, 2016.

[12] M. J. N. O. Panão, N. M. Mateus, and G. C. Graça, "Measured and modeled performance of internal mass as a thermal energy battery for energy fl exible residential buildings," *Applied Energy*, vol. 239, no. October 2018, pp. 252–267, 2019.
[13] H. Johra and P. Heiselberg, "Influence of internal thermal mass on the indoor thermal dynamics and integration of phase change materials in furniture for building energy storage: A review," *Renewable and Sustainable Energy Reviews*, vol. 69, no. September 2015, pp. 19–32, 2017.
[14] H. Hao, B. M. Sanandaji, K. Poolla, and T. L. Vincent, "Aggregate flexibility of thermostatically controlled loads," *IEEE Transactions on Power Systems*, vol. 30, pp. 189–198, Jan 2015.
[15] J. L. Mathieu, M. Kamgarpour, J. Lygeros, G. Andersson, and D. S. Callaway, "Arbitraging intraday wholesale energy market prices with aggregations of thermostatic loads," *IEEE Transactions on Power Systems*, vol. 30, pp. 763–772, March 2015.
[16] J. T. Hughes, A. D. Domnguez-Garca, and K. Poolla, "Identification of virtual battery models for flexible loads," *IEEE Transactions on Power Systems*, vol. 31, pp. 4660–4669, Nov 2016.
[17] A. Ulbig and G. Andersson, "Analyzing operational flexibility of electric power systems," *International Journal of Electrical Power & Energy Systems*, vol. 72, pp. 155–164, 2015.
[18] L. Zhao, W. Zhang, H. Hao, and K. Kalsi, "A geometric approach to aggregate flexibility modeling of thermostatically controlled loads," *IEEE Transactions on Power Systems*, vol. 32, pp. 4721–4731, Nov 2017.
[19] S. Stinner, K. Huchtemann, and D. Müller, "Quantifying the operational flexibility of building energy systems with thermal energy storages," *Applied Energy*, vol. 181, pp. 140 – 154, 2016.
[20] R. G. Junker, A. G. Azar, R. A. Lopes, K. B. Lindberg, G. Reynders, R. Relan, and H. Madsen, "Characterizing the energy flexibility of buildings and districts," *Applied Energy*, vol. 225, pp. 175–182, 2018.
[21] C. Finck, R. Li, R. Kramer, and W. Zeiler, "Quantifying demand flexibility of power-to-heat and thermal energy storage in the control of building heating systems," *Applied Energy*, vol. 209, pp. 409 – 425, 2018.
[22] S. Nolan and M. O'Malley, "Challenges and barriers to demand response deployment and evaluation," *Applied Energy*, vol. 152, pp. 1–10, 2015.
[23] N. Good, "Using behavioural economic theory in modelling of demand response," *Applied Energy*, vol. 239, no. July 2018, pp. 107–116, 2019.
[24] B. Jiang, A. M. Farid, and K. Youcef-Toumi, "Demand side management in a day-ahead wholesale market: A comparison of industrial & social welfare approaches," 2015.
[25] J. Granderson and P. N. Price, "Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models," *Energy*, vol. 66, pp. 981–990, 2014.
[26] K. Coughlin, M. A. Piette, C. Goldman, and S. Kiliccote, "Statistical analysis of baseline load models for non-residential buildings," *Energy and Buildings*, vol. 41, no. 4, pp. 374 – 381, 2009.
[27] T. Walter, P. N. Price, and M. D. Sohn, "Uncertainty estimation improves energy measurement and verification procedures," *Applied Energy*, vol. 130, pp. 230–236, 2014.
[28] R. Sharifi, S. Fathi, and V. Vahidinasab, "Customer baseline load models for residential sector in a smart-grid environment," *Energy Reports*, vol. 2, pp. 74 – 81, 2016.
[29] X. Liang, T. Hong, and G. Qiping, "Improving the accuracy of energy baseline models for commercial buildings with occupancy data," *Applied Energy*, vol. 179, pp. 247–260, 2016.
[30] I. Beil, I. Hiskens, and S. Backhaus, "Round-trip efficiency of fast demand response in a large commercial air conditioner," *Energy and Buildings*, vol. 97, pp. 47 – 55, 2015.
[31] Y. Lin, J. L. Mathieu, J. X. Johnson, I. A. Hiskens, and S. Backhaus, "Explaining inefficiencies in commercial buildings providing power system ancillary services," *Energy and Buildings*, vol. 152, pp. 216 – 226, 2017.
[32] F. Mueller and B. Jansen, "Large-scale demonstration of precise demand response provided by residential heat pumps," *Applied Energy*, vol. 239, pp. 836 – 845, 2019.
[33] Y. Zhang, W. Chen, R. Xu, and J. Black, "A Cluster-Based Method for Calculating Baselines for Residential Loads," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2368–2377, 2016.

[34] F. Wang, S. Member, K. Li, S. Member, C. Liu, Z. Mi, M. Shafie-khah, S. Member, J. P. S. Catalão, and S. Member, "Synchronous Pattern Matching Principle-Based Residential Demand Response Baseline Estimation : Mechanism Analysis and Approach Description," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6972–6985, 2018.

[35] D. H. Vu, K. M. Muttaqi, A. P. Agalgaonkar, and A. Bouzerdoum, "Short-term electricity demand forecasting using autoregressive based time varying model incorporating representative data adjustment," *Applied Energy*, vol. 205, no. August, pp. 790–801, 2017.

[36] Y. Chen, P. Xu, Y. Chu, W. Li, Y. Wu, L. Ni, Y. Bao, and K. Wang, "Short-term electrical load forecasting using the Support Vector Regression (SVR) model to calculate the demand response baseline for office buildings," *Applied Energy*, vol. 195, pp. 659–670, 2017.

[37] "Ecogrid 2.0 [online]." Available: `http://www.ecogrid.dk/en/home_uk`. Accessed: 02/23/2019.

[38] E. M. Larsen, P. Pinson, F. Leimgruber, and F. Judex, "Demand response evaluation and forecasting-methods and results from the ecogrid eu experiment," *Sustainable Energy, Grids and Networks*, vol. 10, pp. 75–83, 2017.

[39] N. O'Connell, P. Pinson, H. Madsen, and M. O'Malley, "Economic dispatch of demand response balancing through asymmetric block offers," *IEEE Transactions on Power Systems*, vol. 31, pp. 2999–3007, July 2016.

[40] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "Stl: A seasonal-trend decomposition," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990.

[41] N. K. Ahmed, A. F. Atiya, N. E. Gayar, and H. El-Shishiny, "An empirical comparison of machine learning models for time series forecasting," *Econometric Reviews*, vol. 29, no. 5-6, pp. 594–621, 2010.

[42] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," *European journal of operational research*, vol. 160, no. 2, pp. 501–514, 2005.