CBS COPENHAGEN BUSINESS SCHOOL
HANDELSHØJSKOLEN

Residential baseline estimation and demand response quantification:

A case study for Czechia

Master Thesis - Copenhagen Business School (CBS)

Authors:
Emil Johan Otto Vesterinen (141294)
Ondřej Beneš (142014)

Supervisors:
Manuel Llorca (CBS - Dept. of Economics)
Martin Pilař (ČEZ a.s.)
Filip Jelínek (ČEZ a.s)

University: Copenhagen Business School (CBS)
Program: MSc. in Advanced Economics and Finance (cand.oecon)
Hand-in date: May 2022
Number of pages: 77
Number of characters: 136,004

Abstract

In the present day volatile electrical energy market, with ever growing demand in combination with decreasing supply due to the need to reduce the usage of fossil fuels, something is needed to balance this possibly occurring gap, in addition to the negative externalities stemming from its existence. This is where demand side management (DSM) would come in. The objective of this thesis is to compare and determine optimal baseline estimation methods inspired by existing literature, to the recently introduced benchmark method set by the Czech regulator ČEPS, as well as quantification of demand response (DR) schemes as a part of the initiative of DSM. The data set for this study, collected over the 2022 heating season, is provided by the Czech energy conglomerate ČEZ a.s as the thesis came to life as part of a pilot project by ČEZ a.s on the potential of DSM in the specific regional setting of Czechia. Therefore, this paper carries a very specific regional setting and framework. We conclude that in the context of residential heat pumps a weather based regression would be the superior model out of the chosen methods in this study, in addition to the accuracy of the baseline method being crucial for the outcome of quantifying the DR. We also note that there are multiple moving parts in addition to kWh usage such as real-time pricing, energy source, and tariffs to take into account for precise DR valuation. Additionally, we suggest that adding the concept of a rebound seems necessary for properly quantifying the value of a DR scheme.

Contents

List of Figures

List of Tables

1 Introduction

In a world that is tending to the become so called "all electric", there is constant pressure on a growing electrical energy demand. Given the nature of daily societal routines which translate into daily energy consumption patterns, this pressure is even more present in so called peak hours. With the additional constraint of phasing out fossil based energy sources, this can prove to be a challenge to energy providers who often guarantee some sort of electricity price range. Having flexible demand which would allow us to balance out our daily load curves by effectively shifting some consumption out of these peak hours and into periods with typically low consumption, may prove to be a vital tool in our transition to a more sustainable and emission free energy sector. However, in order for this demand side management (DSM) to be a viable solution, we need methods for measuring, verifying, and quantifying its potential benefits in addition to identifying its possible drawbacks. This is where baseline estimation becomes the essential piece of a much greater puzzle.

In this paper our main focus, resp. the main research question, is quite clear. After identifying baseline estimation methods usable for our specific data set we test the respective method performances, always comparing them to the benchmark method which arises from our given legal and industrial framework. We compare several methods, both simpler and more sophisticated, while also testing adjustments to these methods. In total, fourteen methods are defined and tested. Additionally, we apply these methods the underlying demand response scheme. Finally, we propose the method which offers the most precise baseline estimation results. We therefore aim to answer the question, which of the methods performs in the best manner and what this means for baseline estimation, resp. demand response valuation. Moreover, we look at the respective traits and limitations of the tested methods.

The framework of this paper is closely related to the region of Czechia. Our context is that of residential heat pumps, the load of which is our main data variable of interest. As this paper is part of a pilot project on residential energy consumption flexibility currently being carried out by the Czech energy conglomerate ČEZ a.s., we work in a very well defined legal and industrial setting. ČEZ a.s. is one of the leading economic entities in the Czech region, operating also in Central and Eastern Europe. With more than 7 million customers it operates in all the activities within the energy sector, acting as producer, distribution system operator (DSO), and retailer. The company is partially state owned

which is why it is also involved in the the talks regarding the implementation of demand response legislation into the current Czech system. Our benchmark method mentioned above and additional methodological specifications are defined by the Czech regulator and transmission system operator (TSO) ČEPS. The data used in this paper has been collected by ČEZ over the duration of the last heating season, that is between December 2021 and March 2022.

The mentioned data set within the ongoing forming of a practically new market in which several new legally defined parties will come to operate is one of the contributions of this paper. The following actual application of chosen baseline methods is what could stand as another new addition to existing literature. Supported by the findings on respective method performance, we offer a first glance at how these methods could work and what their workings in this regional environment are. With Czechia being one of the many European countries looking to eventually rid itself of its dependency on Russian supplied energy, the topic of demand side flexibility will become even more interesting. With this comes the need for an accurate, transparent and effective baseline estimation methodology.

The thesis is structured as follows. In section 2 we define necessary terminology, offer the several motivational viewpoints for DSM and baseline estimation and also offer a snapshot in detail of the current industry status. Section 3 offers a literature review on the tackled issues in this thesis, in addition to several other closely related topics. The data set is meticulously described and presented in section 4. All used baseline methods along with used performance metrics and demand response quantification theory and approaches are covered in section 5. Section 6 presents the results and pitches several discussion topics which are then covered in section 7, along with limitations and other reflections. Finally, section 8 concludes the paper.

2 Industry and Theoretical Background

This section provides a brief description of used terminology, as well as the industry background and the setting in which the paper is written. As noted, the origin of data and region of interest is in this case Czechia. We will start by defining necessary terminology, then present motivation for flexibility (resp. DSM) from several points of view, answer the question why baseline matters and finally explain the role of aggregators and tariffs including a quote on the current status of aggregator and tariff legislation. We also present the baseline codex for balancing services providers as presented by ČEPS (the Czech TSO). This codex defines the paper's main framework as it gives us the benchmark baseline method (Mid 4 of 6) along with general workings of the system. Note that as our main focus and setting is household DR regarding heat pumps (HPs), we do not go into detail on certain topics, e.g. the presence of PV and other sources which turn consumers into so called "prosumers", hence occasionally, we may abstract from some details to keep the paper concise and remain focused on the main topic.

2.1 Defining terminology

The measurement and verification (M&V) generally, and the ***baseline*** more specifically, of demand response (DR) determines the volume of the resource, in our case electricity, and thus is pivotal when determining the value it has to the electric system. Moreover, all customer compensation is based on this M&V and so it has significant influence on the number of types of customers who might be interested in participating in DR programs. There are several ways how to, in detail, define the "baseline" (see section 3), however they all share the main attribute, that is: *"what would the customer's load have been in the absence of a DR event?"*. There are many types of methods how to get the estimation, utilising various inputs and being of various level of complexity. The classification of these methods along with literature examples is presented in section 3, but for now the main principle is sufficient.

Firstly, it is important to clarify the distinction between ***demand side management*** (DSM) and ***demand response*** (DR). The difference between DSM and DR is the focus on load flexibility and short-term customer action (in case of DR) and regular changes in the demand pattern (in case of DSM), as noted by Khabdullin et al. (2017). In other words, DR is a tariff or programme established to incentivise changes in electric con-

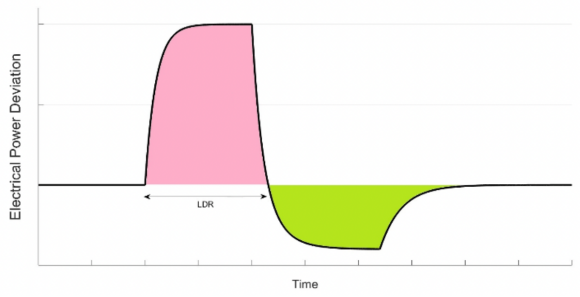
sumption patterns by end-use consumers in response to changes in the price of electricity over time, or to incentivise payments designed to induce lower electricity use at times of high market prices or when grid reliability is jeopardised (Bertoldi et al., 2016). On the other hand, DSM can then be understood as the overall effort to change consumption patterns in the long run which is motivated for example by the aim to allow renewables to supply electricity (energy) at larger than current levels or the overall need for a stable grid (both of which is discussed below in section 2.2). ***Flexibility*** in heat pump (HP) context can then be defined as the possible deviation in electricity consumption pattern of the heat pump (as in e.g. Devriese et al. (2019)). Moreover both Bertoldi et al. (2016) and Bampoulas et al. (2019) divide flexibility (or DR) into two groups:

1. **Explicit** – In this program, demand competes directly with supply in the wholesale, balancing and ancillary services markets through the services of aggregators or single large consumers (more on this in section 2.2). This is achieved through the control of aggregated changes in load traded in electricity markets, providing a comparable resource to generation, and receiving comparable prices. Consumers receive direct payments to change their consumption upon request (to consume more or less). Consumers can earn from their flexibility in electricity consumption individually or by contracting with an aggregator. The latter can either be a third-party aggregator or the customer's retailer (resp. DSO).

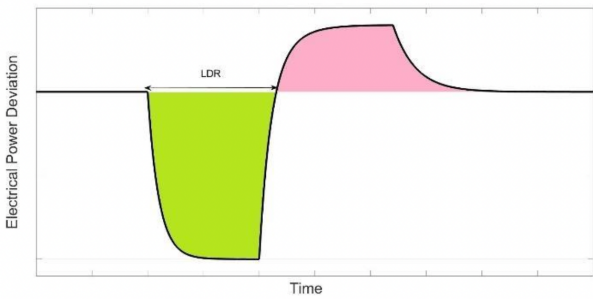
2. **Implicit** – (sometimes called "price-based") refers to consumers choosing to be exposed to time-varying electricity prices or time-varying network tariffs (or both) that partly reflect the value or cost of electricity and/or transportation in different time periods and react to those price differences depending on their own possibilities (no commitment). These prices are always part of their supply contract. Implicit DR does not therefore allow a consumer to participate alongside generation in a market.

We may also classify flexibility based on the direction it takes. This is useful when we talk about the rebound effect below. To clarify, in the case of *upward flexibility* (in HP context), the thermostat set point is increased (by e.g. 1-3°C), and heat is stored passively in the building fabric (see figure 1a). *Downward flexibility* is then the case where, the thermostat set point is decreased (e.g. 1-3 °C), which subsequently decreases the HP power consumption. In this case, heat is curtailed during the modulation period, and it is restored later, for the building to return to the state before the DR action (see the figure

1b). As one might guess, downward flexibility occurs when the aim is load decreasing (e.g. peak shaving), whereas upward flexibility is related to when we aim to "fill out" daily low consumption periods. Figure 1a and 1b illustrate this along with the rebound effect (described below).



(a) Upward flexibility with inverse rebound effect



(b) Downward flexibility with rebound effect

Figure 1: Upward and downward flexibility illustration incl. respective rebound effects.

Source: Bampoulas et al. (2019)

Previously we have mentioned the term "rebound effect", an issue all DR (resp. flexibility) events deal with. In general, the ***rebound effect*** is a term which points out a paradox in energy saving policies, programs and solutions: "*the purchase and use of an energy efficient utility (e.g. car, appliance, etc.) seldom results in the anticipated reduction in the user's energy consumption*" (Klaassen et al., 2017). In the context of flexibility and heat pumps, we have two types of rebound effects: (i) standard rebound effect and (ii) inverse rebound effect. After every downward flexibility (fig. 1b) event, during which the thermostat temperature can fall below desired levels, there is need to reheat the house back to the same desired levels. Therefore, higher than baseline levels of consumption possibly occur after the preceding load reduction, resulting in the standard rebound effect. The rebound effect can also be caused by the people themselves, since one can also take the standpoint where the idea of previously saving energy, justifies above normal-level use in the following periods. The "inverse" rebound effect occurs in the case of upward flexibility (fig. 1a), where the house is usually preheated above desired temperature levels so that there is no need for consumption later during peak hours (a load shift occurs). Following this higher than baseline consumption, it is usually the case that electricity consumption falls below baseline levels, since the thermal mass of the household holds heat and there is no need for active heating. In principle, the effect of the DR event is

only positive if benefits of the load reduction outweigh the extra consumption from the rebound. Various demand response schemes for various appliances and utilities (like HPs, EVs, smart white goods, etc.) have different longitudes of rebound effects. In the context of heat pumps these effects are inevitable, but one should always aim for solutions which minimise the longitude, thus maximising the DR event benefits.

Participation of household customers in DR (resp. DSM), regardless of being only consumers of energy or also as producers (through e.g. PVs, thus becoming "prosumers"), will in every case be pooled, resp. aggregated, and thus some degree of coordination will be required. Parties or bodies who will facilitate this aggregation, "**aggregators**", can either be retailer-based, that is embodied in already established market participants (traditionally distribution system operators (DSOs) and retailers, or stand as an independent party. Aggregators, especially independent ones, have been heavily advocated in recent years and are viewed as a fundamental building block within the European regulation framework which aims to expand the role of demand side flexibility in EU countries (Kerscher and Arboleya, 2022). An EU demand response status report carried out by Bertoldi et al. (2016) defines an aggregator as a service provider who operates – directly or indirectly – a set of demand facilities to sell pools of electric loads as single units in electricity markets. The service is provided separately from any supply contract. As noted before and also in this report, such a service provider may or may not also be a distributor (retailer) of electricity. The status of the presence of independent aggregators for a set of chosen countries is summarised in section 2.3.1.

2.2 Why do we need DSM and why baseline matters?

Demand side management (DSM), demand response (DR) and flexibility are all terms widely used, not only in the EU but in the U.S., Australia, and Japan, when discussing tools on how to make energy cheaper and finding a way how to enable renewable sources of energy (RES) cover larger shares of final consumption. This mainly means levelling out the so called "duck curve"[1] (overall daily consumption curve) so that RES may supply energy without being curtailed as much as possible during low consumption periods, and are able to cover the consumption in full in peak periods. As the volume of installed RES currently isn't able to do so in a consistent and large scale manner, creating a flexible

[1] A reference to the shape the daily consumption curve generally takes, explained e.g. at `https://www.energy.gov/eere/articles/confronting-duck-curve-how-address-over-generation-solar-energy`

environment, where we allow for certain loads to be moved around, may help us in the short-term to cut down on energy produced in traditional, fossil based, plants. The motivations for this are quite clear, as in the 2015 Paris climate conference, the EU and its Member States have committed to limit global warming well below 2°C with reference to pre-industrial levels in the legally binding Paris Agreement (United Nations Climate Change, 2015). In alignment to this the main EU 2030 goals[2] trigger the idea and commitment to use a combination of many small producers of RE, with energy system storage (ESS), electric vehicle (EV) integration and regular power plants infrastructure. As noted by Kerscher and Arboleya (2022), distributed energy sources[3] (DER) and DR coordination result in a potentially bidirectional[4] power flow, which imposes different requirements on the existing infrastructure. Those can be physical, e.g. transmission lines, installed technology etc., and market related, that is mainly in terms of balancing markets, but also for example hedging approaches of DSOs and retailers.

Benefits stemming from an environment with flexible, resp. shiftable consumption at household level are several, however as noted by Kerscher et al., new business models rely on adjusted legal regulations. Nevertheless, they present an opportunity of new value streams and revenues in relation to a changing role of DSOs and TSOs in the market. If done right, each party within the system may benefit. Starting with consumers, for them this brings energy bill minimisation and if households choose to install private (decentralised) energy sources such as PV panels and turn themselves into "prosumers", they can optimise their overall consumption even more, and further capitalise on potentially offering their capacity to suppliers. For retailers, demand response offers means how to hedge their portfolio and for example avoid penalisation[5] from the transmission system operator (TSO) due to caused deviations in the energy balance. Moreover, they suddenly possess a tool to change and tweak their portfolio diagram and so lower their overall costs. With DSM, distributors or distribution system operators (DSO) can, alongside what is noted above for pure retailers, effectively carry out peak shaving, thus pushing down market prices and also lowering their own operation, maintenance and other costs. This may free up funds so desperately needed to invest in the grid expansions and upgrades required to

[2]Include reducing GHG emissions to at least 40% 1990 levels, RES reaching at least 32% of the share in the energy mix and at least 32.5% in energy efficiency improvements.

[3]Include e.g. ESS, PEV, heat pumps, combined heat and power systems (CHP), rooftop PV, etc.

[4]This is not the case of heat pumps and they do not produce any "new" energy. Instead they work as storage, which however still qualifies them as a DER instrument.

[5]Understandably penalisation is the worse of the two outcomes, as parties can also be subsidised if their activity in the (balancing) market with electricity is beneficial to the overall system (grid).

accommodate a future which heavily relies on RES and spread out the projected costs of the overall transition. Finally, TSOs who face the everlasting need to maintain the energy balance and security of the system may benefit from flexibility in the sense that it may bring additional "wiggling room" regarding ancillary services, respectively reserves (FCR, aFRR, mFRR). In the context of heat pumps, the frequency containment reserve (FCR) and frequency restoration reserves (FRR) are, sadly, out of the question since at household level only batteries are capable of such speedy (automatic) responses. Still for manual frequency restoration reserves (mFRR) heat pump flexibility is viable tool how to lower or shift consumption, thus expanding the volume usable for demand-supply balancing. Moreover, for TSOs this means cutting down on other costs associated with energy balancing like the operation of plants used for these purposes. Figure 2 offers a summary of each participants[6] revenues.

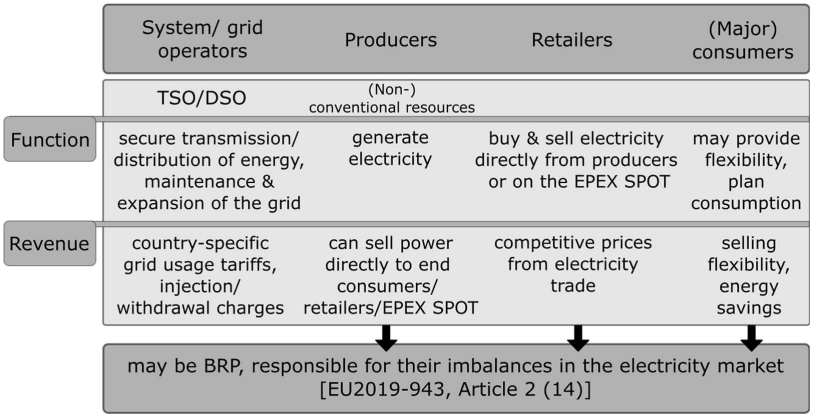


Figure 2: Main traditional stakeholders on the European electricity market, their functions and revenues.

Source: Kerscher and Arboleya (2022)

A final note on how all above relates to baseline estimation, respectively, why does baseline estimation matter. Typically, since demand response programs rely rather heavily on being able to create incentives for energy users to reduce their consumption, there exists a pressing need for a trustworthy, precise and transparent measurement and verification device. Since we are effectively basing the quantification and monetising of benefits on something that did not actually take place (the baseline consumption during the DR event), there must be a degree of trust alongside the legal framework. In other words,

[6]BRP in the figure stands for Balancing Response Party, which can generally refer to any balancing service provider.

the baseline is a "counter-factual," a theoretical measure of what the customer did not do, but would have done, had there not been a DR event (ENERNOC, 2011). This makes baseline estimations a critical component of any DR (thus DSM) program. The baseline is also the primary tool used to measure curtailment during a DR event thus baselines enable grid operators and utilities to measure performance of DR resources. A well-designed baseline benefits all stakeholders by aligning the incentives, actions, and interests of end-user participants, aggregators, utilities, grid operators, and ratepayers.

2.3 Aggregators and tariffs – their role and status

In the context of residential heat pumps, demand response and flexibility ares linked mainly to load shifting, resulting in the levelling out of the daily consumption curve. This is due to the nature of heat[7] demand and it's strong correlation with routines of households. With certain tariffs in place, customer preferences, and technological constraints, all these factors set certain boundaries within which ***aggregators*** will have to operate. Firstly, with growing shares of RES, production in the system a degree of intermittence arises , which inherently requires adequate (additional) means of control. The goal here is very clear, that is using DSM and DR programs to ensure system reliability. As noted in Kerscher and Arboleya (2022), a significant amount of studies have demonstrated a degree of grid congestion and also voltage violation issues[8] in the absence of processes and strategies which would safely integrate high-penetration RES in the current networks. Furthermore, Kerscher et al. point out that prosumers, especially the small ones, who wish to participate in the market, may not possess sufficient capacities for market bidding and may lack the knowledge of market interactions. The integrated or independent aggregator should then be the body that enables this interaction between prosumers and the market. Similarly, the European authorities have defined aggregation as an enabler for prosumer market participation (Council of European Union, 2019). Thus, aggregators potentially hold the power to participate in the electricity market in a beneficial way for various involved parties. Moreover, they stand at the core of facilitating the achievement of climate goals set for the future years and decades. A rather pleasing additional trait of aggregators is that there is room and possible ways for how to incorporate them in

[7]Heat demand both in the sense of heating during the winter season and cooling during summer months as modern day heat pumps are generally capable of providing both.

[8]The voltage frequency for the European continent is 50 Hz. Although there is room for deviation from this frequency, it is extremely small, thus in principle it may be treated as if there was none. Failing to maintain this frequency may result in full black-outs of the given system.

standing European legislation. However, the last does bear several obstacles, which are discussed below in section 2.3.1.

Turning to the so far slightly omitted role of ***tariffs***. To incentivise and activate the flexibility potential of the residential sector (not only heating sector) we must consider certain contract types (tariffs). This gives the understanding of their role in the system some importance. The next several paragraphs offer a short overview of what types of tariffs are currently at our reach. Note that flexible electricity tariffs are nothing new as they have been discussed well over a century, so to summarise, five main contract types for settlement and activation have been defined thus far (Larsen, 2016):

- Volume-based static contracts (e.g. fixed load capping);
- Volume-based dynamic contracts (e.g. dynamic load capping and interruptable contracts);
- Control-based contracts (e.g. direct control);
- Price-based static contracts (e.g. ToU);
- Price-based dynamic contracts (e.g. real-time pricing).

Even though volume-based static contracts are not suitable for household level consumption, this still leaves several suitable candidates. Static price-based contracts, otherwise know as time of use (ToU), are the most popular way of varying final price for consumers. To simplify, they induce people to use electricity during periods when consumption is lower, therefore, prices are set higher during high consumption periods and vice versa (Stamminger and Anstett, 2013). The motivation behind this is to shift some day-time peak consumption to the nighttime lull, effectively increasing the feasibility of conventional generation with long minimum on and off times (such as nuclear), while reducing the need for expensive peaking generators (Larsen, 2016). The consumer then receives a fixed tariff for a longer period of time, typically several months or a year, where prices change two or three times a day.

A trendy tariff related topic are dynamic price-based contracts, a.k.a real time pricing tariffs (RTP). These offer instantaneous pricing of electricity based on the costs of the electricity available for use at the time the electricity is demanded by the customer. Understandably, certain technologies are required (e.g. smart meters, smart heat pumps,

etc.) for this to be possible.

Another type are critical peak pricing tariffs (CPP) which usually combine a ToU and RTP tariff. Here the energy price varies by a time-variable structure with the objective of reducing absolute load peaks at critical times (Stamminger and Anstett, 2013). Inverse CPP tariffs are also possible, where participants are paid for the amounts that they reduce consumption below their predicted consumption levels during critical peak hours. These tariffs are called CPP with a rebate. Stamminger and Ansett also note that several studies on this dynamic pricing offer very different savings results, ranging between almost 0 and 45%.

The tariff style depends on the country and varies from extremes such as 100% energy component based for households in Romania, to 100% fixed and capacity components based for households in the Netherlands, as well as a combination of the two in other countries (Prettico et al., 2019). The next section (2.3.1) briefly looks at the status of aggregators in chosen EU countries, as well as barriers for their existence.

2.3.1 Barriers for Aggregators and their status in the EU context

Even though as stated above there are ways how to incorporate aggregators, in current legislation, established and running frameworks for these independent aggregators are rather rare in the EU. This is in opposition to countries like the US, Australia, South Korea or Japan who have already managed to do so (Bertoldi et al., 2016). This is in part due to certain barriers which we will shortly present. A simple overview and some examples follow.

The final report on demand side flexibility carried out by the EC (European Smart Grids Task Force, 2019) identifies several areas in which some barriers arise. Namely,

- the customer perspective,
- market access,
- flexibility product design
- market process and coordination,
- measurement, validation and settlement of flexibility products,

- technical solutions and platforms to fulfil system and grid needs for flexibility, and
- privacy and security.

Due to the large amount of these topics, we will note only the most relevant ones in the next section.

Firstly, a valid argument is the lack of possible ***standardisation and interoperability*** of the whole system which is very much needed if we talk about user acceptance of DSM. The required presence of certain technologies in households (smart appliances, home energy management solutions, etc.) on the one side, and systems that can incorporate several business cases at once on the side of the energy managers (including peak shaping, facilitating large buildings and industries to provide their flexibility to the grid along with household customers) makes standardisation and interoperability essential. However, as the report states that this condition is currently not fully met and market fragmentation still persists. As one example, we can see different standards for providers of Energy Management Systems (EMS) which in short means having to develop a new device and system for each market. In many cases this might not be worth it, putting customers from different countries at a disadvantage by not giving them access to the same services and opportunities. Furthermore, as also noted in the report, different product design in each country supposes an extra effort and a layer of complexity added to providers that need to adapt their products.

Another topic is ***customer awareness and protection***, which is very relevant in the case of residential customers. As noted not only in the final report, there is a general lack of awareness about what opportunities there are to engage in demand side response. Since understandably customers look toward the financial side of things, this can be attributed to a lack of clear information regarding what is possible (technologically speaking), what is for offer and how well that serves their energy needs, both in terms of their energy bill and additional electricity needs. For example, here the lack of possible comparison of offers may increase risk averseness in potential participants. An obvious data privacy and and security issue stems from this, as does the need for transparent and also high enough financial incentives, which would cover the costs, risks and efforts of setting up the system which allows participation. This is a moment where baseline estimation would come in and offer some hard numbers.

Baseline methodology is also linked to market access issues in terms of ***lack of framework for DR providers***. Clear allocation of energy volumes as well as balancing responsibilities of DR providers are not always present, although it must always be consistent with existing regulation[9]. Also absent is an appropriate, complete, transparent, accurate and standardised methodology for baselining. This is commonly identified as a barrier for access to the market, where especially the lack of transparency may discourage as even slight mistakes in hedging may lead to voluminous financial losses for DSM providers. Moreover, for example lacking clear and unified framework on the above mentioned explicit and implicit DR provision is another topic worth addressing.

Out of those already mentioned, the issues related to measurement, validation and settlement of flexibility products are those probably most closely related to baseline estimations. Finding a baseline methodology that is simple, unbiased, transparent (and all else mentioned above) is far from trivial, especially if we also demand it to be without gaming-options. Inaccurate baselines may render flexibility assets not viable in terms of economical value, thus lowering participation of all parties. The simplicity has been stressed and will be again as it can impact reproducibility, transparency ,and implementation costs, which hurts both customers and providers. Furthermore, gaming-options could easily introduce unwanted effects in day-ahead and intraday markets. Additionally, the concept of independent aggregation typically introduces the need for a Transfer of Energy (ToE). Since ToE has potential to influence wholesale settlement, clear formulations for measurement, valuation and settlement (resp. baseline) are required.

As the above listed topics and legislation in general are not the main topic here, we will not go any further into it here. However, the reason for listing some of the identified barriers is to stress the need for a clear legislative environment, so that it becomes viable for DR providers to enter and operate in the market of flexibility, thus possibly utilising methods which we present and put to work in this paper.

In terms of actual ***status of aggregators*** (bodies utilising baseline methods), Stede et al. (2020) provide us with an up to date overview of companies providing aggregator (enabling) services in Europe. Based on the researched papers and reports and also as noted by Bertoldi et al. (2016), in the EU (resp. Europe), to some degree only Belgium, France, Ireland and the UK have both retailer-based and independent aggregators present

[9]Mainly the Clean Energy Package (CEP) and Electricity Balancing Guideline (EBGL)

in the market. Other countries such as Germany or the Netherlands are in the midst of erasing the withstanding barriers to demand response programs provided by independent aggregators. Other EU countries very often lack frameworks allowing either type of aggregator to even offer such services straight to customers, less so to residential ones. Moreover, from figure 3 we can see that the majority of aggregator services currently focuses on major customers (MC) only, rather than targeting residential customers (RC) as well. Again, Stede et al. (2020) and Lu et al. (2020) provide us with reasoning for why this occurs:

- Existing infrastructures: the diffusion of basic EMS in the industry, for instance, facilitates the identification of DR potentials.
- Forecasting: the baseline electricity consumption in the industry tends to be rather predictable due to commonly planned consumption profiles, and applicable large-scale RES forecasting models exist.
- Data privacy: non-disclosure agreements are already a common practise in industrial collaborations.

Furthermore, Kerscher and Arboleya (2022) note that only a quarter of the companies - those in the role of independent aggregators - do not provide service bundling. The rest do, meaning they combine the roles of the aggregator and for example the role of a balancing response party (BRP). Also, more than half of the aggregator companies directly engage in the wholesale electricity market bidding. Stede et al. (2020) also point out that even though being large enough to participate in wholesale markets themselves, many industrial clients choose to collaborate with an aggregator due to, amongst others, financial hurdles involved with ICT infrastructures.

We leave the topic of aggregators here, and before we move onward to the literature review we will shortly present the codex, respectively framework set up by the Czech TSO, that is ČEPS. The reason for this is the fact that not only does this codex give us our benchmark baseline method which we will always compare to tested methods, but in a more broader sense it also sets the environment we should take into consideration when marking case specific decisions regarding our result interpretation and evaluation, in addition to when making some model related choices.

Company	Ctry.	Aggr./ SaS	Resid./ Major Customer	Functional Roles				
				🔎				
VERBUND AG	AT	ASup	MC	✓	✓	(✓)	✓	✓
Next Kraftwerke	BE	VPP	MC	✓	✓	✓	✓	✓
Hive Power	CH	SaS	RC,MC	✓	✓	✓		✓
BalancePower GmbH	DE	IndepA	MC	✓	✓	✓	✓	
BayWa r.e. GmbH	DE	IndepA	MC	✓	✓	✓	✓	✓
Energy & meteo systems	DE	VPP	MC	✓	✓	(✓)	✓	
GreenCom Networks	DE	SaS	RC,MC	✓	✓	✓	(✓)	✓
gridX	DE	IndepA	RC,MC	✓	✓	✓		
RheinEnergie	DE	ASup	MC	✓	✓	(✓)	✓	✓
Venios GmbH	DE	SaS	MC	✓	✓	(✓)		✓
Plexigrid	ES	SaS	RC,MC	✓	✓	✓		✓
SEAM Group	FI	IndepA	MC	✓	✓	✓	✓	✓
Voltalis	FR	IndepA	RC,MC	✓	✓	✓	✓	
Eneco CrowdNett	NL	ASup	RC,MC	✓	✓	✓	✓	✓
GreenFlux	NL	SaS	MC	✓	✓	✓		✓
ICT	NL	SaS	RC,MC	✓	✓	✓		(✓)
Peeeks BV	NL	SaS	RC	✓	✓	✓		(✓)
Sympower	NL	IndepA	MC	✓	✓	✓	✓	
EmbriQ	NO	SaS	MC	✓	✓	✓		✓
Entelios	NO	ASup	MC	✓	✓	(✓)	✓	✓
eSmart Systems	NO	SaS	MC	✓	✓	✓		✓
GridBeyond	IE/UK	IndepA	MC	✓	✓	✓	✓	
Centrica plc	UK	ASup	MC	✓	✓	✓	✓	✓
Flextricity	UK	ASup	RC,MC	✓	✓	✓	✓	✓
Kaluza	UK	IndepA	RC,MC	✓	✓	✓	✓	✓
Opus One Solutions	UK	SaS	MC	✓	✓	✓		✓
Smarter Grid Solutions	UK	SaS	RC,MC	✓	✓	✓	✓	✓
🔎	Identification of flexibility potential							
	Realisation of potentials							
	Automation							
	Wholsale electricity market participation							
	Bundling of services (aggr./supplier/BRP)							

Figure 3: Companies providing aggregator (enabling) services in Europe assessed according to the functional roles presented by Stede et al. (2020).

ASup – Aggregator Supplier; IndepA - Independent Aggregator; SaS - Software as Service; VPP – Virtual Power Plant.

Source: Kerscher and Arboleya (2022)

2.4 The ČEPS codex for ancillary services providers

In the sections above, especially 2.3.1 we have listed several barriers which limit the large scale deployment of DSM. One of these is the absence of transparent and simple baseline methodologies. The role of the TSO in the Czech context covered by ČEPS, who is the single TSO in the country, unlike in neighbouring Germany for example, which has several TSOs. This year, specifically 16th of March 2022, ČEPS held an online presentation where it introduced a Codex for Providers of Balancing Services (ČEPS, 2022). The motivation arises from the current EU legislation (discussed previously) and the fact that from 2023, ČEPS will be a full time member of the FCRC[10] which is the platform where the cooperation of several TSOs in terms of cross-border FCR procurement takes place. This codex is a positive signal to all market participants, and in the context of this paper an important piece of documentation. We will mention the most relevant parts, that is those that affect our main focus.

Firstly, baseline is presented as the tool which will be used to measure and control the balancing services provided, thus replacing previous methodology. This allows market entrance for a wider pool of providers and participants, most importantly, it allows participants from the demand side to actively enter the balancing market. It is also noted that this baseline methodology applies to energy utilities of the so called 2nd category, that is those which have either an installed capacity smaller than 30 MW, or where the voltage level of their connection node is under 110 kV. This very clearly targets small providers, resp. calls for aggregation of small capacity participants. Each unit in the market will also be allowed to simultaneously provide multiple services using the baseline methodology, thus allowing for the presence of aggregators.

As for us only mFRR services are relevant, since heat pumps aren't capable to supply FCR and aFRR, we will note the calculation approaches related to mFFR, which for the ***provider***[11] are,

- The prediction interval is to be 30 minutes with a one second frequency (e.g. one sent at 00:00:01 applies until 00:30:01).

[10] Information about the workings of this platform as well as results of it's functioning are available at `https://www.regelleistung.net/ext/`

[11] Only those truly relevant are listed.

- The prediction interval starts 8 minutes before the mFRR supply interval and ends 5 minutes after this interval.
- The MAPE value will be calculated at 1 hour intervals from minute values.

and for the ***TSO*** (ČEPS) they are:

- The baseline method will be the general method of Mid 4 of 6.
- The baseline method will be adjusted by the multiplicative adjustment method.

The codex also lists that for the evaluation of accuracy of predictions within the provision interval, the mean absolute percentage error (MAPE) will be used as formulated and described in section 5.5.

3 Literature review

In this section a literature review is provided. While the main focus lies with actual baseline and rebound estimation papers, we aim to include papers on the wider scope of DSM and DR research and aggregators as well, since the field itself is only a piece in a much larger puzzle. In literature, customer baseline load (CBL), or simply baseline, is defined in several ways. Wi et al. (2009) defines it as the "usual" pattern of electric demand, respectively it tells the consumed load when no interference takes place. In our context, that is household heat pumps, this would be consumption of the heat pump under no flexibility occurrence. Sun et al. (2019b) use the terminology of dynamic time of use events (dToU) and non-time-of-use events (non-ToU). The former are periods with low or high price events, where a dToU group is a group of customers who have received an experimental (meaning flexible or other than default) dToU tariff. The latter are periods where default prices apply to the customers in such a non-ToU group. To quantify the DR event we can take historical values or estimations of the demand change between the dToU and non-ToU (baseline) cases. Lastly, we could also turn to Zhang et al. (2016) and use the term load reduction which is defined as the CBL subtracted by the actual load. This distinction is offered here in order to illustrate that when understanding baseline results, one should always bear in mind the legislative and terminology context in which these results have been generated. Still, the main idea stands clear, baseline is what would have been consumed if no DR events took place.

3.1 Baseline & Rebound

As one would expect, baseline and rebound estimation methodology literature closely tracks publications on DR and DSM topics. Even though these baseline estimations of energy consumption are usually only a part of bigger DSM solutions, they are the building block for the evaluation and measuring of DR and should be of utmost concern for policy makers and DR program designers. It comes as no surprise then, that the literature on various estimation techniques is vast. If we understand baseline (and rebound) estimation as described in section 2, we can see a trend in the sophistication and complexity of these methods. However, as pointed out in e.g. Mohajeryami et al. (2016) simplicity is an important trait when designing and choosing baseline methods. The following text presents a brief overview of published methodology on baseline estimation, however one last note is in order. If we were to take the general grouping of baseline methods as

presented in the 2011 white paper by ENERNOC (ENERNOC, 2011), in this paper we focus only on *Baseline Type I* methods, hence such methods where baseline is generated using historical interval meter data and potentially also using weather and/or historical load data to generate a profile baseline. Other groups may include *maximum base load, meter-before meter-after, baseline type II* and *generation*, but these are not considered here[12] as those methods are inferior.

As described in section 5, when observing a methodology "name" one should always distinguish between the method's historical load selection mechanism and the actual estimation methodology. For example, if we take the High X of Y method, the "high X of Y" refers to the selection rule, as the method uses the highest X days (in terms of total load) of the last Y acceptable days for estimation. The actual estimation though, is a simple averaging method. Usually, when grouping methods, the actual estimation approach is used and we will follow this trend. Thus, based on the literature reviewed for this paper, we feel it suitable to group covered methods as follows:

- averaging methods,
- regression methods,
- control group or cluster methods and
- other advanced methods.

The following text very briefly summarises the main characteristics of each method, whereas table 1 offers an overview of some papers on baseline (rebound) estimation which have been published as part of an underlying DSM trial or project.

Averaging methods are methods which use short historical data of close to given date loads and calculate CBL by averaging the load of previous non-event days (days when no DR occurred). These methods typically first choose reference days[13] (denoted usually as Y) and then from these a number of similar non-event days (denoted usually as X) are chosen for the actual averaging. Taking our context of 15 minute interval data, this would mean we always take the average from all chosen X days for each 15 minute interval as our baseline. The values of Y and X differ from application to application, as do the

[12]Each of the listed methods is described in the ENERNOC white paper, page 6.

[13]Throughout the text, these are often referred to as "acceptable days".

selection criteria resulting in a variety of methods, as pointed out by Lee (2019). For example, the New York Independent System Operator (NYISO) High 5 of 10 method (Jazaeri et al. (2016), where high translates to ordering the Y reference days by highest total daily load and taking the highest X (here 5) days for averaging[14]. Another example would be the California ISO (CAISO) which used High 10 of 10 (Wang and Tang, 2020). Other possibilities to the "high" selection criterion are "mid" and "low", where mid refers to taking the middle X (dropping the top and bottom days) and low the lowest X from the Y reference days. The Mid 4 of 6 method will be the benchmark in this paper as it will be used by the Czech TSO (ČEPS). It is also possible to use other selection mechanisms for averaging, e.g. a moving average based approach, used by e.g. the New England ISO (ISONE) and probably to most advanced manner of baseline estimation within the averaging methods.

Regression-based methods use regression models to "forecast" the baseline load patterns during event periods. The difficulty here is to find the most accurate non-linear relationship between chosen features (regressors) such as temperature, historical load, etc., and baseline load (the dependent variable here). In opposition to simple averaging, resp. similar-day methods, regression methods can be capable of including event day information in the model itself. However, as Sun et al. (2019a) point out, since DR events are usually established for extreme weather conditions, there arises and issue of insufficient historical measurements in order for one to build a reliable model of this type. This calls for the use of adjustments, which is further explained in section 5 and discussed in section 7.

One way of overcoming the issues above can be using load profiles of customers who have not been subject to any DR event but who match the typical load profile (TLP) of the subject who's CLP we are estimating. Such methods, **control group (cluster-based) methods**, use various clustering methods (e.g. K-means) to link DR event customers to clusters (control groups) of similar non-event customers, thus overcoming challenges such as limited data during the extreme days Hatton et al. (2016). The main caveat of this method type is the clustering itself. In order for these models to be effective and precise, finding optimal matches of load profiles is key (Sun et al., 2019a). In the absence of this historical data, synchronous pattern matching has been proposed by e.g. Wang et al.

[14]The listed methods in this sentence apply to weekday DR events. For weekend DR events the methods are often different, e.g. the NYISO use the High 2 of 3 method for such occurrences.

(2018) as an effective solution.

Other advanced methods include the use of more sophisticated tools such as neural networks combined with machine learning, handling the stochastic uncertainty (which is key for probabilistic estimation) and more. Some authors, e.g. Sun et al. (2019a) even try to exploit both pre-event (historical) and post-event ("future") dependencies. Probabilistic estimation is another well explored group of methods. Vallés et al. (2018) use probabilistic quantile regression, Weng et al. (2018) explore the possibilities of probabilistic Gaussian process regressions and so on. Some authors also combine cluster-based methods with regression method, such as Sun et al. (2019b) who combine deep embedded clustering with quantile forest regression. Generally, these advanced methods give more precise results to the previous methods, however, there are some drawbacks. The most noteworthy is the need for massive datasets of fine-grained data and also the computational time of some of these approaches may become an issue if time is of the essence. Moreover as noted before, such methods may not be suitable when evaluating the simplicity of the baseline estimation as it would be probably safe to assume that most DR participants, especially in the context of households, may not be fully comfortable with terms like Gaussian processes, deep learning, neural networks, etc.

Even though sophisticated methods often produce better results, due to mentioned simplicity requirements and also necessary computational time for example, they seem to be more suitable for, e.g. post-DSM evaluation when the only true factor is the most precise evaluation of the DR. This is also why regulators most often turn to averaging methods. To increase their ability to compete with other methods, **adjustments** can offer additional precision. Some sort of adjustment can complement all estimation methods, regardless the type. Generally, there are two approaches. The more simpler manner is an *additive* adjustment, where one simply adds the difference between the baseline estimation and actual load to the baseline estimated value at the beginning of the DR event. In other words, we observe how "off" the baseline estimate is at the start of the DR event and add this error to the baseline itself throughout the whole DR event to adjust for any event day change in load. This adjustment can also take multiple periods prior to the DR event start and use their average differences as the final baseline adjustment (as in ISONE Grimm (2008)). The alternative, a *multiplicative* adjustment, is similar with one main difference, that is, the adjustment is done by multiplying the baseline by a calculated "adjustment multiplicative" which is obtained again from hours prior to the DR event.

The two mechanisms are described in more detail in the methodology section (section 5). Both adjustment methods are very often used along with similar-day (X of Y) methods as these methods give high estimation errors particularly due to neglecting event day data as in Mohajeryami et al. (2016). Based on the papers which include (and compare) multiple methods, we can expect decreasing error sizes with the increase in sophistication of the methods.

Table 1: Relevant Baseline Estimation Papers (chronological order)

Paper	**Estimation method**	**Data & Notes**
Haas and Biermayr (2000)	Regression based.	Space heating data, Austria 1970-1995.
Wi et al. (2009)	Regression (exponential smoothing) with weather adjustment.	Past load data, PJM 2001-2003.
Wijaya et al. (2014)	Averaging methods (Low X of Y, etc.).	Irish CER smart metering trial dataset. Includes High(Mid) XofY and regression as well.
Klaassen et al. (2017)	Multiple regression and artificial neural networks.	Dutch PowerMatching City (PMC) Pilot data.
Jazaeri et al. (2016)	Averaging, regression, machine learning and polynomial interpolation methods.	Comparison paper. Smart meter data from Victoria, Australia.
Zhang et al. (2016)	Cluster based with morning adjustments.	Data form residential customers in a city in Southern U.S.
Hatton et al. (2016)	Control (cluster) based.	Une Bretagne Avance trial DR with 280 DR participants and a control group of 433.
Mohajeryami et al. (2016)	Averaging methods and regression, including adjustments.	Irish CER smart metering trial dataset, here 262 customers from 2009.
Weng et al. (2018)	Probabilistic via Gaussian regression.	PG&E and OhmConnect datasets from year 2011-2012 and 2014 respectively.
Vallés et al. (2018)	Probabilistic quantile regression.	Data from a DR field test in Castellón de la Plana (ESP), part of EU ADDRESS project.
Wang et al. (2018)	Cluster based - variations of Synchronous Pattern Matching.	Irish CER smart metering trial dataset (2010).
Sun et al. (2019b)	Cluster based (DEC) plus quantile regression forests.	Low Carbon London Progam data.
Lee (2019)	Averaging methods (variations of Mid XofY).	South Korean (DRTM) and French (NEBEFDR) mechanisms data.
Müller and Jansen (2019)	Probabilistic estimation	Data from 300 residential buildings with heap pumps.
Sun et al. (2019a)	Probabilistic based (Bayesian Deep Bidirectional LSTM Neural Network training).	Low Carbon London Program data.
Wang et al. (2019)	Probabilistic based, combining (averaging) quantile regression methods.	ISONE and Irish CER smart metering trial data between 2013 and 2016.
Mancini et al. (2020)	Cluster based (household archetypes, resp. dwelling clustering)	–

3.2 Other relevant literature

3.2.1 DSM and DR

Since demand side management (DSM) and DR trials are rather large and complex projects, the same goes for the literature that aims to present them. A good example of this would be the multi-phase the PowerMatching City Pilot in the Netherlands, which has, for several years, been the focus of a range of papers from network (market) modelling (e.g. Bliek et al. (2010)), including smart grid coordination (e.g. Kok (2013)) and multi-goal system optimisation (e.g. Wijbenga et al. (2014) to those closest to our focus, that is baseline estimation (Klaassen et al. (2017). The scale, area (region), included technologies (e.g. heat pumps, EVs, photovoltaics[15], etc.), present tariffs, characteristics of buildings (apartment buildings, residential houses, industrial complexes, etc.) and also ones standpoint (profit-seeking company, regulator, policymaker, etc.) are all determinants in how these projects end up being set up, modelled and evaluated. As this paper's main focus is not the modelling and general set up of the underlying DR scheme, and moreover, the size of the studied group is only 27 household heat pumps with no other utilities included, we will only touch the main topics associated with DSM and DR in general.

Smart grid modelling is a increasingly growing topic of it's own, where demand side load management is only one of the three domestic technologies usable, the other two being distributed generation (e.g. PV) and energy storage. A good example that the interest in this field is nothing new is the 2010 paper by Molderink et al. (2010) where the authors offer an overview and general concept of then current research, along with some promising results of field tests. As one of the main motivations for DSM is to allow for higher penetration of renewables into the energy mix[16], models which can evaluate this are also needed. Luckily such models have been around for some time now, as in Pina et al. (2012). A good example of DSM modelling in a heat pump specific context would be e.g. Hedegaard and Balyk (2013) where authors present a model that facilitates analysing individual heat pumps and complementing heat storage in integration with the energy system, while optimising both investments and operation. The results lead authors to

[15]The presence of any individual energy source, usually PV but possibly even EVs, adds an important new element to any DSM since it turns consumers into so called "prosumers". Thus, households interact not only as the demand side, but carry certain traits of the energy supply side too. The main being their ability to reserve and sell their production capacity to their distributor for e.g. balancing purposes. Here the pricing and set up of DSM becomes even more complex, as showed e.g. in Venizelou et al. (2020).

[16]Mainly by levelling out the daily consumption curve, the so called "duck curve".

note the potential which heat pumps may have in terms of integrating wind power at a national level, including impacts on investments, system costs, fuel consumption, and CO_2 emissions.

Another interesting question is the willingness of households to participate in DR in general. This is tackled by for example Srivastava et al. (2020) who operate with a data set of 186 respondents and focus on Belgium winter electricity peaks where the main question relates to the willingness to accept limits on the use of home appliances in return for a compensation. They find that willingness to enrol in a program increases with age, environmental consciousness, home ownership, and lower privacy concerns. Moreover, the analysis predicts that 95% of the sample surveyed could enrol in a daily load control program for a compensation of €41 per household per year, and the authors summarise that while an initial roll out among older and more pro-environment homeowners could be successful, a wider implementation would require an explanation of its environmental and financial benefits to the population, and a greater consideration of their data privacy concerns.

As mentioned above, buildings themselves are also an often concern as generally, insulation and other characteristics have great power in reducing consumed energy. In the context of the Danish newly built buildings regulations Foteinaki et al. (2018) study the differences between a single-family house and an apartment block in terms thermal storage capacity and what are the effects on flexibility. The findings showed that low-energy buildings are highly robust and can remain autonomous for several hours and that the potential for storage in the thermal mass is considerable. They also state that the analysis presented high dependence of flexibility potential on boundary conditions (ambient temperature, solar radiation, internal gains) and underlined the importance of envelope insulation. Similarly, Le Dréau and Heiselberg (2016) also study the effects of short term heat storage in thermal mass on the energy flexibility of residential buildings.

Scale is also mentioned above as a determinant in DSM and DR methodology approaches. As one might expect the approach mainly differs between urban scale DSM projects, as in e.g. Hedegaard et al. (2019) where the authors include a whole residential neighbourhood in Aarhus, Denmark, or simply taking a single representative as in Bampoulas et al. (2019) who take a household type representing about 40% of the Irish building stock and infer conclusions based on this. However, ultimately, these two approaches obviously share

their aims. Also, whenever the scale grows, the optimisation issues gain on relevancy as one has to handle different preferences, constraints and habits. Jin et al. (2017) offer a good example of the development of a micro-grid optimal dispatch with DR (MOD-DR) as they demonstrate a 17.5% peak load reduction and 8.8% cost savings on a campus prototype.

3.2.2 Aggregators

The idea of aggregating small residential loads and allowing them to be bid to an open (liberated) market is the principle motivation for any DSM or DR solution. This principle and the theoretical background for aggregators was described in section 2.3 and so below we only provide literature support and potential further readings. Probably the most recent study looking at the role of aggregators in the EU legislation and energy framework is the one by Kerscher and Arboleya (2022). In this study the authors present a magnificent techno-economic review of aggregator models and case studies[17] and agree with the EU regulation that assigns aggregators (especially independent ones) a pivotal role in the upcoming energy transformation. However, they do confirm the cited findings in section 2 and point out that aggregators face various regulatory, technical, and economic barriers. A paper analysing the German balancing mechanism by Koliou et al. (2014) illustrates that DR has indeed potential, but is undermined by three mechanism design aspects: minimum bidding volume, minimum bid duration and binding up and down bids. For a consolidated insight into the optimisation functioning and bidding methodology expected of the debated aggregators one can look to the paper by Babar et al. (2017), for example. As mentioned, in this paper the "prosumer" role of households is not considered, however, Brusco et al. (2014) study the workings of an energy district as an aggregated coalition of prosumers in the context of a Italian residential housing area. Finally, for completeness, as we work only with household consumption, we abstract from industrial demand response, however, for example Stede et al. (2020) offer a rather recent case study on industrial DR in a German context.

[17] A tabulated overview of listed models and studies is available in Table 3 of the paper – `https://www.sciencedirect.com/science/article/pii/S0142061521006001?via%3Dihub#t0015`

4 Data

This section presents the used data set, how it was obtained and handled. Furthermore it describes the workings of the underlying demand response scheme and finally offers basic descriptive statistics.

4.1 Data Origin, Collection and Handling

The data used in this paper has been collected by the energy conglomerate ČEZ a.s. as part of a DR and flexibility pilot project during the 2021/2022 winter heating season. To filter out possible sources of bias and other misleading factors, households from a single region and county were targeted. As a result, all of the thirty-one households, which were willing to take part and had the technical necessities installed[18], come from the Domažlice county located in the Pilsen region, south-west Bohemia in Czechia.

Usable data is available for the period starting on the 11$^{\text{th}}$ of December 2021 and ending on the 26$^{\text{th}}$ of March 2022[19]. This presents us with almost four full months (105 days) of observations. As mentioned above, the original sample group contained 31 households, however, data on four heat pumps (HPs) had to be omitted due to excessive amounts of NA values. The presence of such values stems from the technical setup of data logging. To cut down on data transfer volumes, only changes in heat pump load were logged. This means the intervals between logs could be anything between several seconds to tens of minutes, depending on momentarily consumption levels. This data was then internally processed by ČEZ so that the final output would be minute data with the condition, that in the case of any NA value, the last known (logged) load value was prolonged for a maximum of 20 minutes. Beyond this 20 minute limit, the NA values were left as NAs, as there was some uncertainty about why the HP wasn't recording any values. Most probably this was due to technical reasons, e.g. connection issues. The outside temperature was logged similarly, with the difference that the last known value was prolonged up to the next logged temperature, as we can be certain about the presence of outside temperature. All collected variables are described in detail in section 4.3.

[18]Those being an electric or hybrid heat pump (HP) and a smart meter. Moreover, to be able to effectively collect data, ideally the same company would be the heat pump supplier. In this case Tenaur s.r.o., which is also part of ČEZ Group.

[19]The sample was cut here to avoid complications related to the time change.

Since simply omitting the NA values would result in an incomplete time series, and negatively effect, or even restrain, our attempts at obtaining the baseline estimation, a way how to replace them had to be found. Here we decided on the following approach, which was also the recommended manner by ČEZ. For each NA value we took the average load of all the other heat pumps in the sample for that time period[20]. Given the fact we had two "types" of HPs, those undergoing DR events and those not, for each "type" we used load from heat pumps with a matching state of consumption. Thus, if the HP is consuming in a DR decrease request state for example, only loads of HPs also in the state of decreased consumption were used to obtain the NA replacement value for that period. This means that those HPs which did not undergo any DR event throughout the whole sample time[21] were not included in this calculation in all of the DR event periods. Similarly, we couldn't use data from HPs undergoing DR events when replacing NA values for heat pumps which didn't undergo any interventions. The DR mechanism properties are discussed in full in section 4.2.

Finally, since the current EU regulation tackling imbalance settlement harmonisation adopts the rule the *"by 1st January 2021, the imbalance settlement period must be 15 minutes in all scheduling areas, unless regulatory authorities have granted a derogation or an exemption (derogations may be granted only until 31 December 2024)*[22]" (Council of European Union, 2019), and because imbalance settlements are one of the main motivation factors for this paper (as discussed in section 2.2), we transformed the initial minute data into 15 minute intervals. This was done by taking the average load for each HP for the given 15 minutes. Therefore, from the original time series of thirty-one heat pumps with minute frequency data, we have a dataset with 27 heatpumps with 15 minute frequency, resulting in 274,752 observations of heatpump load[23]. Out of these 22 HPs underwent DR events and 5 did not. Table 2 summarises.

[20]If there were NAs also in the other HP load data for the given period, these were dropped from the calculation. We faced one occurrence of full NA row (row 144,584). This row was replaced with the average loads of ± 30 rows.

[21]These serve as the control group.

[22]Furthermore, the Article 8(4) also states that "from 1 January 2025, the imbalance settlement period must not exceed 30 minutes where an exemption has been granted by all the regulatory authorities within a synchronous area".

[23]If we include the remaining three variables – date-time, requests and outside temperature – we get 305,280 observations.

Table 2: Number of Observations per Heat Pump Type

	DR event HPs	non-DR event HPs
# of HPs per type	22	5
# observations at 15 min interval	10,176	10,176
total observations per type	223,872	50,880

4.2 The Underlying Demand Response Mechanism

To provide full context on the understanding of the data sample and also the results presented later, a final note on the underlying DR event scheme is necessary.

As described above in table 5, request values could be of either the value 0, 1 or 2. Zero corresponds to no intervention into the heat pump load, and so for periods in which this value is present in the data, the state of consumption is identical for both groups of HPs, both controlled and uncontrolled. In those periods where this value equals 1, a request to *decrease* load was sent to the heat pump, thus resulting in a downward DR event, where the HP expectedly consumes below the later estimated baseline. Similarly for values of 2, in which case the difference is that an *increase* request was sent, hence the HP expectedly consumes above baseline. Furthermore, as pointed out be e.g. Wijaya et al. (2014), two types of signals are possible for residential customers:

1. A signal which communicates the DR event start/end times and the amount of kWh to be reduced.
2. A signal which communicates the DR event start/end times and lets the customer decide how much she is willing to reduce.

Our specific underlying DR scheme uses signals of the second type, therefore it is to the customer to choose the volume by which he or she wishes to decrease (increase) the load. This type of DR signal has some interesting implications in regards to customer incentives and possibilities in "gaming" the system. These are discussed in section 7.

Table 3 shows the amount of requests along with type, for the 15 minute interval data[24]. As clearly visible, the DR scheme was more active in December, January and March, whereas less active in February. Since each request relates to a 15 minute interval, the

[24]Recall that initial data was in 1 minute frequency. For this data, the requests count obviously differ.

total sum of hours, in which the HP load were intervened with sum up to 322 hours (or just over 13 days), for all the given months.

Table 3: DR Requests Count

Request type	December 2021	January 2022	February 2022	March 2022	Total
No request (0)	1677	2473	2664	2074	8888
Decrease request (1)	210	314	8	258	790
Increase request (2)	129	189	16	164	498
Controlling requests (1 & 2)	339	503	24	422	1288

In total, we have 68 and 46 days on which a decrease or increase event takes place, respectively. These dates are available in the appendix (tables 15 and 16). The DR requests are timed on a weekly basis where each day always follows a given time schedule of both decrease and increase requests. Table 4 offers an overview of how the requests were timed every week. Note, that most of the days we have two of each request type, except for Thursday and Sunday. Also, Wednesday is rather peculiar, as its first DR event of the day is a decreasing one, followed by a increasing event. Also, unlike in the other days, on Wednesday, the two requests are always consecutive, which can pose some difficulties when setting up the DR valuation, e.g. when wanting to quantify the rebound. Requests occur in all months although in February, this is true only the first day of the month[25].

Table 4: DR scheme — weekly time schedule

Weekdays	1st Increase req. (2)	1st Decrease req. (1)	2nd Increase req. (2)	2nd Decrease req. (1)
Monday	01:00-02:00	06:00-07:00	11:00-12:00	17:00-18:00
Tuesday	01:00-03:00	06:00-08:00	11:00-13:00	17:00-19:00
Wednesday	11:00-13:00	10:00-11:00	16:30-17:00	17:00-19:00
Thursday	–	06:00-06:30	–	17:00-19:00
Friday	–	–	–	–
Saturday	01:00-02:00	09:00-11:00	14:00-15:15	18:00-19:00
Sunday	–	09:00-11:00	–	17:00-19:00

[25]Even though one might find oneself questioning the size of the dataset, we would like to note, that based on the literature review of similar papers, often only a handful (or even a single) of DR events are studied, as the main focus is not the DSM scheme, but the performance of each method. The same applies to our case. In our case, maybe the opposite could be the issue, that is, having enough non-DR days which we can actually use for baseline estimation. Luckily, having a full month where almost no DR events occur makes for a large enough control sample.

Here it should be noted that the DR scheme itself is not the main focus of this paper, as apparent from the lack of some real-time factors[26] when deciding the timing and scale of the DR requests. It does, however, follow the general principle of peak-shaving and "filling in" the low consumption periods of the day. This is clear from the request time intervals described above. As the core focus of this paper is the comparison and evaluation of the chosen baseline estimation methods, we will abstract from any judgement of the DR scheme itself. The main output will be a proposal on which methods are most likely to generate the lowest errors in our regional and time interval (15 minute) setting.

4.3 Variable Descriptions

This section provides a detailed overview of all the variables obtained and used. Table 5 offers a short description of each data variable, including the the ID of this variable used in code and also throughout the text (mostly in section 5). Detailed descriptions of each variable follow below, followed by relevant descriptive statistics.

Table 5: Data Variables – Basic Description

The *id* represents variable classes. DT is used as *POSIXct*, out_temp as *numeric*, req as *integer* and hp(i)_j as *numeric*. The expression $j = \{t, f\}$ and represents the type of HP, that is controlled (t) and uncontrolled (f).

Variable	Simple description	id
Date-time stamp	Date and time expressed in the ISO 8601 format.	*DT*
Outside temperature	Outside temperate (°C) calculated as the average temperature of all household locations included in the pilot.	*out_temp*
DR request	DR request signal with values 0 (no request), 1 (decrease request) and 2 (increase request).	*req*
Heat pump load	Actual load of household i heat pump in kW.	*hp(i)_j*

Our ***date-time*** variable is reported in the ISO 8601 format[27] and as noted already, ranges from the 11$^{\text{th}}$ of December 2021 to the 26$^{\text{th}}$ of March 2022. The ***outside temperature*** (in °C) variable is collected in order to have some representative of weather when necessary. Since all the locations (households) of the heat pumps were within the same region, the final values are computed as the average temperature of all the locations at any given time

[26] E.g. real-time-pricing as discussed in section 2.

[27] That is: YYYY-MM-DD HH:MM:SS.

period. There was no distinction between controlled (undergoing DR events) and uncontrolled (not undergoing DR events) heat pumps in the case of temperature. Regarding the ***DR request*** variable, as noted above, being a request variable, it can take the value of 0, 1 and 2. Each relating to a DR request type, "no request", "decrease load request" and "increase load request", respectively. The DR scheme itself is described in detail in section 4.2. Finally, the ***heat pump load*** is recorded for a total of 27 heat pumps. As noted in table 2, 22 heat pumps undergo load interventions, whereas 5 heat pumps do not and serve as a control group. The originally minute frequency data underwent transformation into data with a 15 minute interval, for reasons described in section 4.1, by taking the average load of all the minutes between e.g. 00:01 and 00:15 for the final value at 00:15. Table 6 offers an overview of basic descriptive statistics for all variables. As it seemed excessive, and also fairly irrelevant, to report these statistics for each individual heat pump, the HPs were grouped by type (controlled and uncontrolled) and the average of the given statistic for all the individual HPs is reported (columns *"all hp_t"* and *"all hp_f"* in table 6 below).

Table 6: Basic Descriptive Statistics

Statistic	DT	out_temp	all hp_t	all hp_f
Min.:	2022-01-01 00:00:00	-6.1044	0.0000	0.0000
1st Qu.:	–	0.1561	0.5608	0.6653
Median:	–	2.1911	1.6969	2.1960
Mean:	–	2.6835	1.6992	2.0236
3rd Qu.:	–	4.8115	2.6118	3.0368
Max.:	2022-02-28 23:45:00	20.5681	6.0923	7.7318

Before starting with the actual estimations, in order to understand the data, several visualisations might be helpful. First, we grouped the controlled HPs by request type and took the average load of all HPs for each time interval within a day. The output of this is shown in figure 4, 5 and 6 for the request types 0, 1 and 2, respectively.

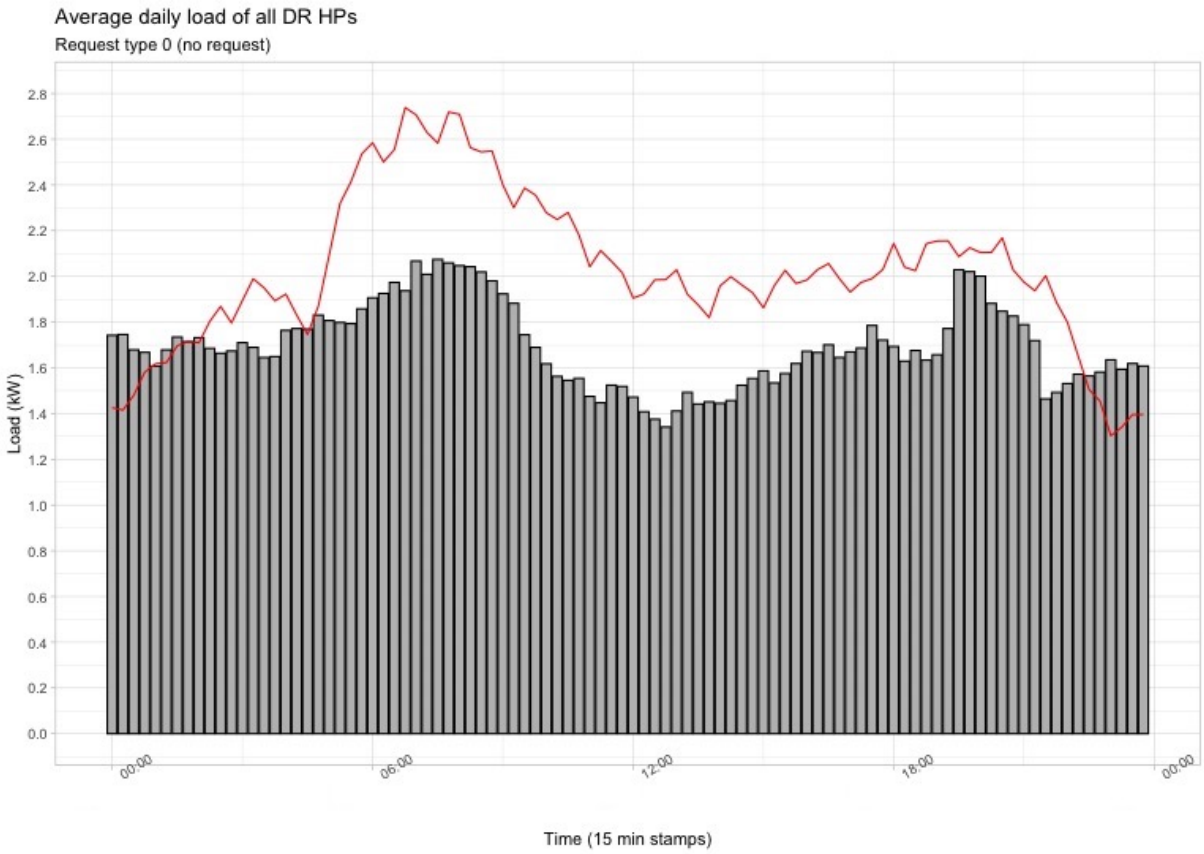


Figure 4: Average daily load of all DR HPs – request type 0 (no request)

The red line in figure 4 represents the average load of all the 5 uncontrolled heat pumps. It is included only here, as these are the days and hours in which controlled HPs were not following any requests (reg = 0), thus their load were not intervened with.

Looking at figure 4 we should keep in mind that the effect of February is present here. The February days are the non-DR event days in the data sample. If we were to generate this plot only for February (not reported here), we see the exact same daily seasonality, only with slightly lower values (around 0.2 kWh) which may be explained by lower temperatures in January and December relative to February (see fig. 8). The relationship applies the March, however, here it is inverse. From the plot we can see the daily peaks, between 08:00 and 09:00, and between 18:00 and 19:00. The daily lows during night time and around noon are also rather clear. The average maximum load (peaks) are somewhere just above 2 kWh and the daily minimum goes as low as just under 1.4 kWh. The red line in figure 4 represents the average daily load for our group of uncontrolled HPs. Interestingly, although this group follows a more or less similar daily consumption curve, this group has slightly higher load. Here the small size of the group may be the explanation, as taking an average from 5 HPs may be more prone to affects of extreme

values (outliers). This difference persists even for when plotting just February data (not reported here).

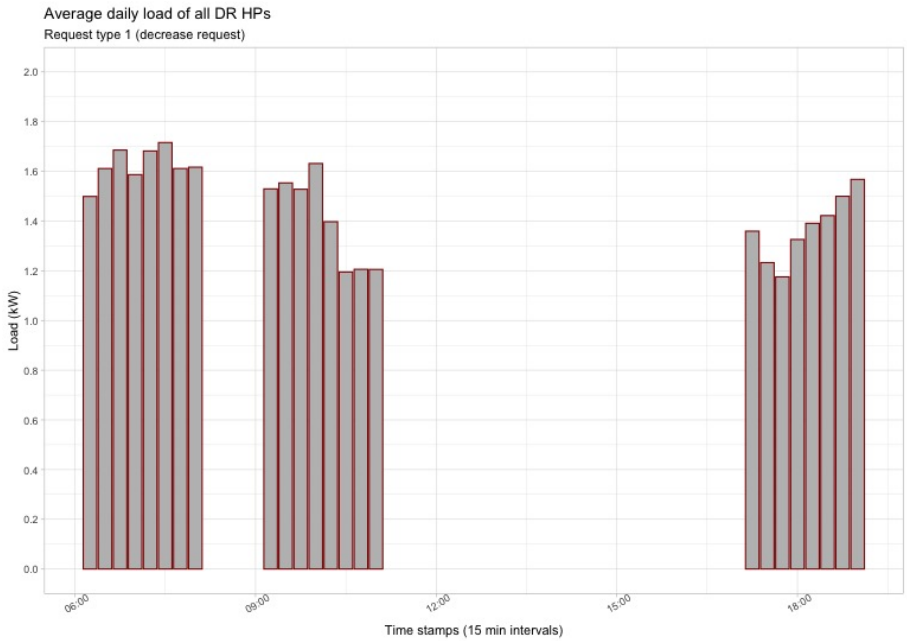


Figure 5: Average daily load of all DR HPs – request type 1 (decrease request)

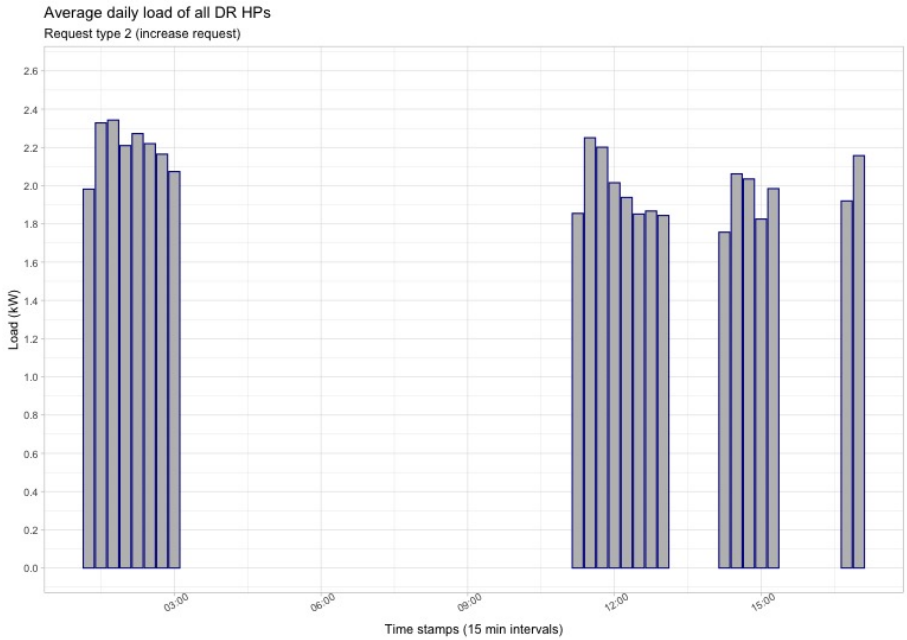


Figure 6: Average daily load of all DR HPs – request type 2 (increase request)

Figure 5 shows the average load during decreasing DR events as described in table 4. Recall that the decreasing request time intervals were not the same for each day, thus

resulting in the gaps during the morning interval. The decreased load drops as far as below 1.2 kWh in some cases. Also, in general the average load doesn't exceed 1.7 kWh even in peak hours. If we were to compare the load levels to those in fig. 4 we can see that there is clearly some decrease in consumption. These moments are those which will interest us the most in our baseline estimations. Finally, figure 6 shows the average load of increase request time intervals. Again, we can observe a difference between these values and those in figure 4 for given time stamps, only here as expected the load increases, in some moments, even as far as beyond 2.3 kWh[28].

Another useful visual description is to see the summed consumption of our controlled HPs for the four months. Figure 7 plots total controlled HP load in for December, January, February and March at 15 minute intervals. As we already know how the daily seasonality looks like, the small interval doesn't matter here. Instead, here we are interested in the general level of consumption. Clearly, consumption was higher in December and January, most probably as a result of lower outside temperature, as reported in figure 8. When observing the two graphs, we can see a clear inverse relationship in general.

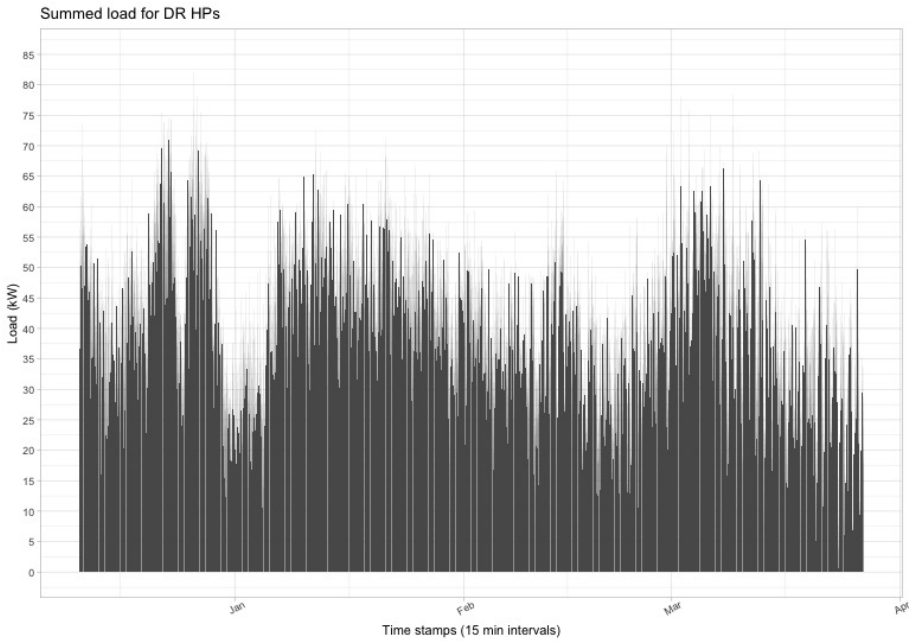


Figure 7: Summed load of all DR HPs (15min intervals)

[28]Note that these values are the average values of all the controlled HPs for the given time stamp, thus the individual load increase can (and often did) go beyond even 2.8 kWh.

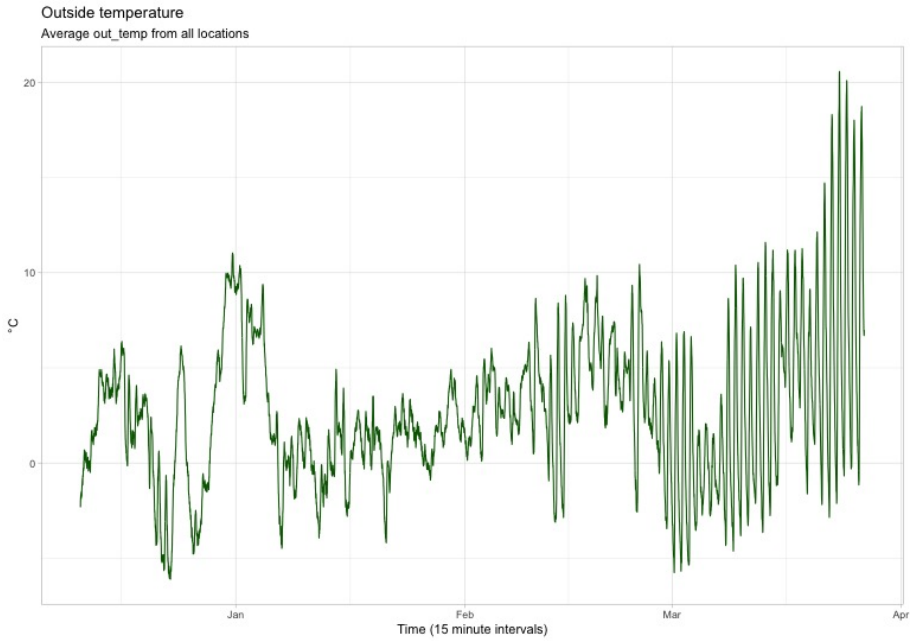


Figure 8: Average outside temperature for all HP locations

5 Methodology

In this section we describe all used baseline estimations models. We also present the carried out adjustment method and how the DR events, including the rebound where quantified. At this point it should be stated that based on the classification of ENERNOC (2011) we consider only so called Baseline Type I estimation methods. In these methods the baseline is generated using historical interval meter data and may also use weather and/or historical load data to generate a profile baseline that usually changes hour-by-hour[29], which is shown to be the most prominent. Furthermore, we would like to stress that even though no estimate is perfect, there are some methods superior to others or best suited to specific programs or customer types (ENERNOC, 2011). When valuating a baseline method many sources, e.g. Mohajeryami et al. (2016) or ENERNOC (2011), define three factors that are critical above all others. Those being: accuracy, simplicity and integrity, and one should always keep these three in mind when choosing or building a method. A final note, all coding in this paper was done using the RStudio software and the code will be made available on GitHub. The section continues with the simple averaging methods, then moves to exponential moving average and exponential smoothing and finally presents the two regression models. It then moves to presenting the used performance metrics and the approaches and methods utilised to quantify the DR scheme along with the rebound.

5.1 Simple Averaging Methods

If we abstract from the simplest of baseline estimation methods, that is the *last Y days* method, averaging methods with the various X of Y selection rules could be considered as the more simplistic approach to baseline estimation. However, they are used by many (independent) system operators (ISO) as they offer a very straightforward, transparent and easy to understand[30] methodology, all of which are traits that have been noted in many papers (see section 3) to be crucial when deciding a good baseline method.

The general approach of these methods is to select a group of most recent days from

[29]Other baseline types include: Maximum Base Load, Meter Before – Meter After, Baseline Type II and Generation.

[30]Especially for the customer, who's willingness to take part in the DR program can very well be diminished by a complicated method which he or she are unable to understand and thus possibly feel lack of trust. This is covered in section 3) and also the discussed in section 7.

the set of *"acceptable days"* based on several conditions listed below, and then compute the average for each time slot using a subset of these days. In other words, they share a selection rule, last X of Y, and they all have the same estimation method, averaging. What they differ in is the way they select the X days from the last Y days for the actual baseline estimation. The following sections present each of the three estimated X of Y methods: High X of Y, Low X of Y and Mid X of Y.

5.1.1 High 5 of 10

As mentioned, X of Y methods and especially the high X of Y method, is rather popular amongst ISOs, e.g. in the United States. Papers by Wang and Tang (2020) or Wijaya et al. (2014) confirm this offering the following :

- PJM: High 4 of 5 for a weekday, and High 2 of 3 for a weekend DR event.
- NYISO: High 5 of 10 for a weekday, and High 2 of 3 for a weekend DR event.
- CAISO: High 10 of 10 for a weekday, and High 4 of 4 for a weekend DR event.

The main reasoning being that most DR events[31] happen on days where the usage, and especially peak usage, are expected to be high. Thus, having a method able to replicate these peak consumptions for DR event hours is desirable and most representative of the actual potential consumption. For programs working all year round, especially programs using both decrease and increase requests, this may not always be the case. As a result some operators opt for the Middle X of Y (ENERNOC, 2011), which is what ČEPS has published as their used method (ČEPS, 2022). This method is described later on. As several specifics of the High X of Y method differ from application to application, we choose to base our approach on the findings of Mohajeryami et al. (2016), Wang and Tang (2020) and Grimm (2008).

High X of Y methods are defined by Jazaeri et al. (2016) as follows: From the original pool of the last Z calendar days, the last Y working days are selected after applying the exclusion rules. Here we follow the general approach and exclude all DR days, holidays and also weekends when estimating for workdays. One could also generate more sophisticated exclusion rules and exclude for example outlier days, however, this can decrease the set

[31] Many papers handle only decreasing DR events. In this paper we always have a combination of multiple decreasing and increasing events, often in one single day. See section 4.2.

of acceptable days and so we do not apply this rule. Next, the daily load (the sum of loads from each individual time slot) of each of those Y days is calculated. The Y days are ranked according to their daily load from the highest to the lowest and the highest X days are selected. The baseline load of the event day is then the average of the same time slot load from all of the X days. Mohajeryami et al. (2016) describes a widely used high X of Y method used by the New York ISO. This method includes the following steps:

1. First, Y non-DR days must be selected. The NYISO method chooses Y = 10. Apart from pervious DR event day, weekends and holidays are also excluded.

2. Next, X days are chosen from the aforesaid Y days based on the level of consumption. The NYISO has X = 5 and chooses the highest consumption days.

3. Finally, taking the average load of these five days gives us the baseline.

If high X of Y is defined as:

$$High(X, Y, d) \subseteq D(Y, d),$$

then the high X of Y baseline of customer i $\subseteq$ C for time slot t $\subseteq$ T on event day d is the following:

$$b_i(d, t) = \frac{1}{X} \cdot \sum_{d \in High(X,Y,d)} l_i(d, t) \tag{1}$$

The NYISO also uses an additional algorithm to select the past 10 non-DR days, which they use for their "Day-Ahead Demand Response Program". The goal of this is to choose the 5 days with highest kWh usage from a pool of 10 days that meet the requirements of the algorithm, instead of them just being the past 10 non-DR days as e.g. in the California ISO method. This algorithm is described in more detail in Grimm (2008). However, due to limitations in data size, we choose not to add additional exclusion rules so that we preserve as much data as possible and thus choose to use the simpler approach of the CAISO and simply use the last 10 acceptable days. Therefore, in summary, we take our X and Y from the NYISO but abstract from the additional condition of the X days, as in the CAISO.

Table 7 offers a snapshot from our X of Y methods selection mechanism. Note that the DR days in this case are workdays between the 7^{th} and 10^{th} of March thus the vector $D(Y, d)$ is filled with all the days in column *Date*. Once the X days are identified, the

baseline can be obtained as shown in table 8.

Table 7: Selection rule mechanism for each "X of Y" method

The High 5 of 10, Low 5 of 10 and Mid 4 of 6 are used for illustrating how the respective X of Y methods choose the days used for baseline estimation. This is why the *mid 4 of 6* column contains only four x values.

Date	hp0_t	high 5 of 10	mid 4 of 6	low 5 of 10
2022-02-16	162.46	x_5		
2022-02-17	129.22			x_1
2022-02-18	138.05			x_2
2022-02-21	171.09	x_4		
2022-02-22	173.47	x_3	x_2	
2022-02-23	153.10		x_4	x_4
2022-02-24	148.18			x_3
2022-02-25	162.03		x_3	x_5
2022-02-28	213.20	x_1		
2022-03-04	210.46	x_2	x_1	

Table 8: Baseline calculation example for the High 5of10 method

Time	2022-02-16	2022-02-21	2022-02-22	2022-03-04	2022-02-28	baseline hp0_t
00:00:00	2.67	1.09	1.46	2.61	0.93	1.75
00:15:00	2.43	1.70	1.57	2.42	3.68	2.36
00:30:00	0.15	1.92	1.68	2.60	1.69	1.61
00:45:00	0.00	2.08	2.19	1.23	4.07	1.92
01:00:00	0.91	1.99	1.91	0.00	3.35	1.63
01:15:00	1.96	1.73	3.21	2.02	2.76	2.34
01:30:00	3.28	1.73	2.56	4.16	0.26	2.40
01:45:00	3.23	1.00	2.26	3.38	0.00	1.97
⋮	⋮	⋮	⋮	⋮	⋮	⋮
22:00:00	0.57	2.53	1.29	4.32	2.50	2.24
22:15:00	0.00	2.12	1.59	3.44	0.29	1.49
22:30:00	0.00	0.61	1.91	3.32	0.00	1.17
22:45:00	1.57	0.00	2.34	1.89	1.07	1.37
23:00:00	2.03	0.14	2.00	0.51	4.17	1.77
23:15:00	2.03	2.74	1.22	0.00	3.39	1.88
23:30:00	1.98	2.87	1.27	0.18	3.13	1.88
23:45:00	1.80	2.55	1.15	0.00	2.25	1.55

5.1.2 Low 5 of 10

Mohajeryami et al. (2016) define low XofY as well as it basically mirrors the high XofY. The methods are essentially the same with the distinction that the former chooses the X days with the lowest daily consumption, instead of the highest consumption days. The reasons for this and implications which stem from this are discussed later in the results. As a result, if low X of Y is defined as

$$Low(X, Y, d) \subseteq D(Y, d),$$

then the low X of Y baseline of customer i $\subseteq$ C for timeslot t $\subseteq$ T on day d is the following:

$$b_i(d, t) = \frac{1}{X} \times \sum_{d \in Low(X,Y,d)} l_i(d, t). \tag{2}$$

Once again, if we look to table 7 we can compare which days would be chosen in comparison with the other two methods. The baseline calculation is then identical to table 8, obviously using different days, thus different loads.

5.1.3 Mid 4 of 6 – The Benchmark Method

Middle (or Mid) X of Y is again very similar to the two previous methods, however carries one main difference as it chooses the "middle" X of the previous Y days which means it drops the extreme values first and uses the remaining days for estimation. With Mid X of Y defined as:

$$Mid(X, Y, d) \subseteq D(Y, d),$$

the mid X of Y baseline of customer i $\subseteq$ C for time slot t $\subseteq$ T on day d is then:

$$b_i(d, t) = \frac{1}{X} \times \sum_{d \in Mid(X,Y,d)} l_i(d, t). \tag{3}$$

Given the selection mechanism dropping always at least the one highest and one lowest Y day, there is always a difference of at least two between Y and X. Here one could raise the question why we do not use the same length of Y for all methods. Given the framework of this thesis (see section 2) we choose the method used by ČEPS to be our benchmark method to which we compare the remaining methods. It could also be argued that it would then make sense to use the same Y values, that is six, for all the other X of Y

methods. Here again our argument stands. As we are very clearly trying to identify a somewhat better performing alternative to the regulators chosen method, there is little reason to limit all our remaining proposed methods simply to match regulators choices when in reality companies will not be required to do so. We therefore stick to our choices of Y being 10 and X being 5 for the high and low X of Y.

5.2 Exponential Moving Average

Although also being an averaging method, the exponential moving average method is considered to be somewhat more sophisticated than the methods above. Also, compared to the other averaging methods it has one possible advantage since it is able to weight days closer the actual DR event day more strongly. In this paper we choose utilise two moving average methods. The first is noted as "exponential moving average" (EMA) while the second as "exponential smoothing". Both methods are rather similar, however inspired by different papers and also they handle certain specifics of the estimation slightly differently. The main distinction is the moving average equation terms set up, the way they set their respective weights and also how they handle the calculation of the initial load. Both are described below.

5.2.1 EMA

The use of exponential moving average is very well described in Mohajeryami et al. (2016). The method is basically a weighted average of customers' historical data from the beginning of their subscription. We would like to stress that only days d from the set of acceptable days D enter the calculations. The method begins with computing an initial average load of a customer. Then, continues with calculating an exponential moving average using this initial average load. The baseline for the customer is achieved at the end.

The mathematical representation can be defined as follows. Let $D(\infty, d) = \{d_1, ..., d_k\}$ also $1 \leq \tau \leq k$ to be constant. This constant is the number of days used to determine $EMA_i(d_\tau, t)$ (eq. 4) which is the initial average load for customer $i \subseteq C$ or time slots $t \subseteq T$. Here it is important to stress that the τ chosen days are days prior to the day in which the customer joins the DR program. Note the difference in indexes, which are τ for the initial load days and j for the DR program days.

$$EMA_i(d_\tau, t) = \frac{1}{\tau} \sum_{j=1}^{\tau} l_i(d_j, t) \tag{4}$$

The exponential moving average for $\tau \leq j \leq k$ is then

$$EMA_i(d_j, t) = \delta \cdot EMA_i(d_{j-1}, t) + (1 - \delta) \cdot l_i(d_j, t) \tag{5}$$

where $\delta \in [0, 1]$. The choice of δ and τ is up to method user. The weight of each day decreases exponentially with time, which is what theoretically makes this method superior to the simple averaging methods. The baseline for customer $i \subseteq C$ on day d for time slot $t \subseteq T$ can then be calculated as follows:

$$b_i(d, t) = EMA_i(d_k, t) \tag{6}$$

An obvious drawback of this method if defined as above, is that there is no way how to include event day load data, as the baseline is technically the last preceding day's estimated EMA for every time slot. This can later be handled using adjustments. Adjustments are explained in section 5.4 below. It should also be noted that for days earlier than $d_{\tau+1}$ the baseline cannot be defined in this method. This means that if a DR event happens during this short time, the baseline for this customer cannot be calculated. This results in DR programs not being able to include these customers in the program until they get access to enough days for them to be able to calculate the initial average load.

One ISO that uses this methodology is the New England ISO (ISONE). ISONE employs the following algorithm. For new customers of the DR program, the hourly average of the five ($\tau = 5$) previous business days is used as the initial load. Similarly to other ISOs, weekends, holidays and other event days are excluded from this calculation. This initial load is referred to as the CBL 6, where the six represents the 6^{th} day after the five business days. The equation for calculating CBL 6 is as follows:

$$CBL_6 = \frac{\sum_{i=1}^{5} l_i(d, t)}{5}$$

After the CBL 6 has been calculated, the new customer can be referred to as a current customer, which means that equation 4 with $\delta = 0.9$ can be used to calculate the next day's baseline. The EMA is calculated for every following acceptable day (again excluding weekends, DR days and holidays). Having $\delta = 0.9$ essentially means that a 90% weight is given on the previous days CBL and a 10% weight is given to the consumption of the current day. As our EMA method we choose to use the same contants as the ISONE, that is δ equal to 0.9 and τ equal to 5.

Unfortunately, given the nature of our data where DR events occur from the very beginning of the data sample, we do not have the five business day loads which are needed to calculate the CBL 6. We handle this by taking the first five acceptable days for estimation and use these as input into the CBL 6 formula. In reality this means that our vector D of acceptable days must be shortened not only for this one method, but also for all the other methods. Specifically, the days in table 17 which also appear in table 9 have to be dropped, resp. treated as if they are days preceding DR program entry. This is the closest we can get to the method of say, ISONE. Nevertheless, as observable in table 9, the days are more than a month back from DR event[32]. One could technically use more days to obtain the intial load, but since we want to follow the method of ISONE as closely as possible for reference, we leave τ equal to five.

Table 9: Chosen days for EMA initial load calculation

#	Date	Weekday
1	2021-12-17	Fri
2	2021-12-24	Fri
3	2021-12-31	Fri
4	2022-01-07	Fri
5	2022-01-14	Fri

A final note. Given the nature of our data, all of the days used to calculate initial load are Fridays, which is definitely something to note. But as mentioned earlier, this is our only option in order to follow ISONE as closely as possible.

[32]Recall that the DR event days for which we estimate the baseline are days between the 7^{th} and 10^{th} of March 2022.

5.2.2 Exponential Smoothing

When using this method we build on a well cited paper by Wi et al. (2009). As noted above, they way Wi et al. sets up the exponential smoothing equation slightly differs from eq. 5. As with the EMA, the formula in eq. 7 is used to calculate the moving average from the first estimated day to the day preceding the DR event day. Even though inspired by Wi et al., we opt to make one slight change in the formula below. Instead of using the time slot $t-1$ as Wi et al. do, we use the time slot t as we don't see a good reason why one would calculate the given day moving average using the estimation from not only one day back, but also one time slot back. To us, using the same time slot seems like a more reasonable approach. In actual values, this means we estimate the load for e.g. 12:00 using load from one day back at 12:00, whereas Wi et al. would, based on their formula, use the load from 11:00 (for an hourly data frequency). Using the $t-1$ time slot before would make sense if we were using same day data, however, this is not the case as we are using previous day estimation and load values. The formula is thus as follows:

$$b_i(d_j, t) = \alpha \cdot l_i(d_{j-1}, t) + (1 - \alpha) \cdot b_i(d_{j-1}, t), \tag{7}$$

where $l_i(d_{j-1}, t)$ is the actual consumption in time slot t on day $d-1$ and $b_i(d-1, t)$ is the baseline estimate (resp. the calculated exponential smoothed average) for time slot t for day $d-1$. As in the EMA approach also here the baseline is then given by the estimated values for the last preceding acceptable day.

The exponential smoothing method differs from the EMA in two more ways. Secondly, it doesn't calculate the CBL_6 or any other averaged initial load to start the model, but use the first day's load as the estimate and get the model rolling from the second day. This means we lose one observation (day). Thirdly, the weights (α) are not chosen to be constant but instead they are defined using an exponential formulation, like so:

$$\alpha = \frac{2}{N+1}$$

Given this definition, the weights change over time, as N changes, resp. grows. This means that the closer we are to the event day, the larger is the weight of the second term in eq. 7, that is the estimated moving average, resp. the baseline.

One can clearly see that both the EMA and the exponential smoothing methods are unable to include event day data. Wi et al handle this using an adjustment, however, as such an adjustments can be made to all the listed methods, we define these separately in section 5.4. A final note, the exponential smoothing stated like so could also be viewed as an autoregressive integrated moving average (ARIMA) (0,1,1) without a constant. This stems from having a simple AR(1) model (random walk) and correcting this model so that instead of forecasting simply based on the most recent value, it filters out noise be using an average of the last few observations, thus effectively correcting the local mean estimate. Therefore, we can write:

$$b_t = b_{t-1} + \alpha \cdot \varepsilon_{t-1}$$

where $\alpha \cdot \varepsilon$ is the (exponentially) weighted error term and because $\varepsilon_{t-1} = l_{t-1} - b_{t-1}$ by definition, we can write

$$b_t = b_{t-1} + \alpha \cdot (l_{t-1} - b_{t-1}),$$

and by rearranging we get

$$b_t = \alpha \cdot l_{t-1} + (1 - \alpha) \cdot b_{t-1}.$$

Which results into an ARIMA (0,1,1) without a constant. Note that in terms of notation, b_t being our baseline is in fact an estimation of the load, thus could also be notes as $\hat{l}_t$, which is used in section 5.3 when specifying the model. This is also why Wi et al. refer to this method as a regression method.

5.3 Linear and Quadratic Regression

Regression models are widely proposed in baseline estimation methodology, as they offer a way how to incorporate additional variables, which in the context of heat pumps, is very relevant since e.g. weather obviously affects the need for heating and cooling, thus affecting heat pump consumption. Linear regressions are, therefore, a commonly used method for calculating customer baseline. Wi et al. (2009) state that the reason why most of regressions used for estimation are linear is due to simplicity. Mohajeryami et al. (2016) defines a possible regression for baseline estimation for customer $i \subseteq C$ on day d for time slot $t \subseteq T$ is as follows:

$$b_i(d,t) = (\beta_{it})^T x_{it} + \varepsilon_{it} \tag{8}$$

where the feature vector x_{it} for this econometric model can include many different variables connected to consumption, like outside temperature, humidity, historical load and many more. Since for us the only available data other than load is outside temperature (see section 4), we choose to work with only this additional explanatory variable.

The general algorithm of baseline estimation using regression goes as follows.

1. Identify the set of acceptable days for estimation (D).
2. Run the regression model (see below) for each time slot (in our case 96) using all the observations in D for that given time slot.
3. Estimate the event day baseline using coefficients from step 2.

In this paper we choose to use the ordinary least squares (OLS) estimator (model) to obtain the coefficients. Also, in order to capture a potential parabolic relationship outside temperature has with electricity consumption on heat pumps, our regressions will include outside temperature and additionally outside temperature squared. When looking at electricity consumption as a combination of many household utilities, finding a parabolic relationship with outside temperature isn't necessarily as obvious, but since we are doing this specifically for heat pumps that are able to both heat up and cool down the living space, it makes sense to use a quadratic regression as well. As shown in the figure below, there are indicators for a parabolic relationship existing.

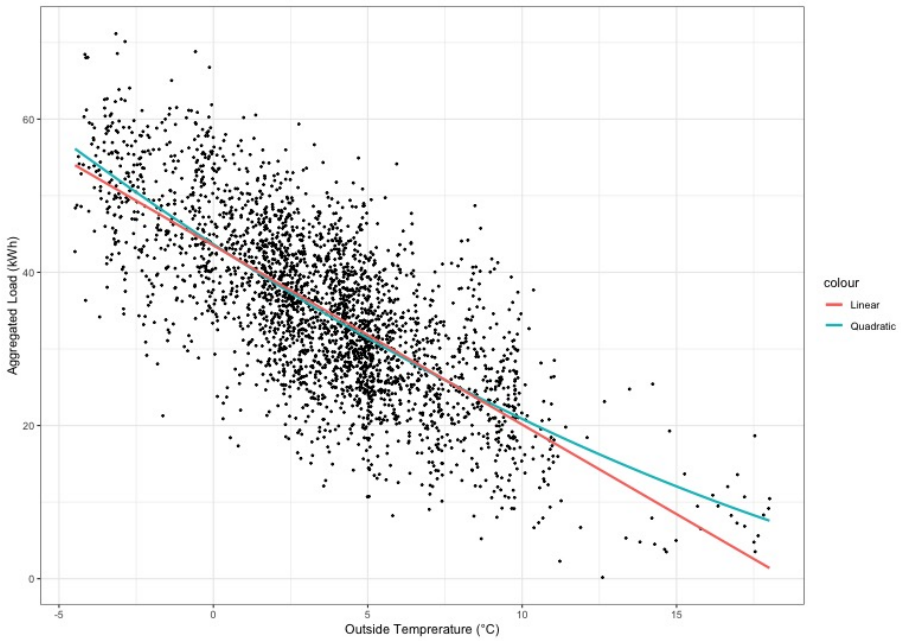


Figure 9: Load grouped by date and time slot regressed on outside temperature.

One of the reasons why the parabolic relationship isn't as prominent as it could be is because our data has been obtained during the pure heating season (December to early spring), where temperature-wise we would still be at the downward slope of the parabola. Even though our maximum temperature in the data sample is approximately 20.5 °C, the density for "cooling" temperature range is very low as we can see in figure 9. We believe that when the outside temperature moves beyond twenty in a more permanent manner, we will be moving towards the upwards slope of the parabola. Additionally, when doing these regressions for each heat pump and each time slot separately, parabolic curves were often the results of plotting and statistical testing[33].

This means that the OLS[34] regressions we will be using for baseline estimations are the following:

$$\widehat{l_{it}} = \alpha + \beta_1 \cdot out\_temp_{it} + \varepsilon_{it} \tag{9}$$

$$\widehat{l_{it}} = \delta + \gamma_1 \cdot out\_temp_{it} + \gamma_2 \cdot out\_temp_{it}^2 + \varepsilon_{it} \tag{10}$$

[33]The results of statistical testing are not reported in the paper as the focus of this paper is not their statistical significance but instead the capability of the method to estimate the baseline. For completeness, the t-statistics and p-values rarely render the coefficients for the squared temperature terms statistically significant at 5% level, however, often they are significant at 10 % level.

[34]We base our OLS theoretical framework on Wooldridge (2020)

Making the equations for baseline estimation:

$$b_i(d,t) = \hat{\alpha} + \hat{\beta}_1 \cdot out\_temp_{idt} \tag{11}$$

and

$$b_i(d,t) = \hat{\delta} + \hat{\gamma_1} \cdot out\_temp_{idt} + \hat{\gamma_2} \cdot out\_temp_{idt}^2 \tag{12}$$

respectively. Note that the dependent variable here is the load, and the explanatory variable are the outside temperature terms. Also, for notation purposes we separate $\hat{l}_{it}$ and b_{it} even though for the DR event day, they are the same thing. This is done to show the plugging in of the estimated coefficients from equations 9 and 10 into the baseline equations as $\hat{\beta}$ and $\hat{\gamma}$.

This specification means that we run one linear and one quadratic OLS regression for the relationship between actual load and outside temperature for all preceding acceptable days, separately for all time slots. The coefficients obtained from these regressions are then used to calculate the baseline for the DR_event day based on its actual outside temperature at each time slot. In real world application these regressions would be rerun every time additional acceptable days have been obtained. This way the most updated coefficients are used for calculation.

A final comment regarding possible regression models. Although considered, weighting our observations, resp. opting for a GLS model wouldn't necessarily be better for this purpose since in our case the temperature should be as important regardless of what observation it is, as it correlates so closely with heat pump electricity consumption. When doing this for complete household electricity usage one might want to consider using GLS not only to weight the observations time wise but also differentiate between utilities, since then the outside temperature doesn't carry as much explanatory power as if we consider solely heat pumps. In our data, the small heteroscedasticity found, could be explained by the fact of having very few observations temperatures over 10 °C. The carried statistical testing (available in the appendix, section 8) indicates that heteroscedasticity isn't that big of an issue here.

5.4 Adjustments

Most of the methods stated above, especially the X of Y (similar-day) methods, by definition fail to include the event day information (load), thus rendering high errors. The reason this is a problem could for example be weather conditions changing during the event day, causing the load to shift significantly in one direction. In order to test this in our dataset, we will include results of each method adjusted for event day load changes. The load reduction during event periods is then defined as the difference between the adjusted baseline and the actual load instead of the original baseline. The adjustment is defined by a time frame. This time frame changes depending on application, but is normally around 2–4 time slots before the start of the event.

Adjustments to each method are often seen as an integral part in making a baseline estimate more accurate (Mohajeryami et al., 2016) since as shown above, many ISOs use X of Y methods which call for an adjustment. When making adjustments, the two most commonly found methods are the additive and the multiplicative methods. A multiplicative adjustment uses the percentage change and applies it to the estimated baseline. An additive adjustment utilises the absolute change.

Grimm (2008) goes through the process how ISONE makes adjustments to exponential moving average using the additive method. ISONE makes an adjustment to the baseline only on an event day by calculating the average difference between the new baseline (the baseline calculated with $\delta = 0.9$) and the actual customer load two hours prior to the event period. The adjustment is then applied to the two hours prior, in addition to all the hours during the event. This adjustment is made only if the new baseline is lower than the actual load at the start of the event.

The other, more commonly used method is the multiplicative method. Grimm (2008) goes through the example of how the PJM Interchange Energy Market ELRP makes multiplicative adjustments to its estimations. The multiplicative method works by first taking the average of actual load and the average of estimated baseline two and three time slots prior to the DR event. Then, the average load from these two periods is divided by the average estimated baseline from the same periods in order to get a percentage difference, the *"adjustment multiplicative"* (AM) in eq. 13. The adjusted baseline is then the product of the original baseline and the AM for time slots within the DR event duration. The AM formula can be written as:

$$AM = \frac{Average(l_i(d, t-2) : l_i(d, t-3))}{Average(b_i(d, t-2) : b_i(d, t-3))} \tag{13}$$

where $l_i(d, t-2)$ is the actual consumption at time $t-2$ on day d whereas ($b_i(d, t-2)$ is the predicted consumption at time $t-2$ on day d before the start of the DR event. An example of how the AM is applied is shown in table 10.

Table 10: Multiplicative adjustment example for DR events 10:00–12:45 and 16:30–18:45

Time	Actual load	Baseline	AM	Adjusted Baseline
0.00.00	3.58	1.75		1.75
⋮	⋮	⋮		⋮
9.15.00	1.65	2.23		2.23
9.30.00	1.39	1.47		1.47
9.45.00	1.86	1.68		1.68
10.00.00	4.05	2.33	0.82	1.91
10.15.00	1.01	2.12	0.82	1.74
10.30.00	1.01	2.13	0.82	1.75
10.45.00	0.58	2.57	0.82	2.10
11.00.00	0.56	2.42	0.82	1.99
11.15.00	1.28	2.24	0.82	1.84
11.30.00	2.48	2.17	0.82	1.78
11.45.00	2.56	2.03	0.82	1.67
12.00.00	2.55	1.40	0.82	1.14
12.15.00	2.09	1.77	0.82	1.45
12.30.00	1.91	1.92	0.82	1.58
12.45.00	2.43	2.52	0.82	2.07
13.00.00	2.23	1.73		1.73
⋮	⋮	⋮		⋮
15.45.00	1.08	2.10		2.10
16.00.00	1.99	2.32		2.32
16.15.00	2.23	1.97		1.97
16.30.00	2.17	1.83	0.69	1.27
16.45.00	2.07	2.01	0.69	1.40
17.00.00	2.10	1.99	0.69	1.38
17.15.00	1.20	1.81	0.69	1.26
17.30.00	0.86	1.76	0.69	1.22
17.45.00	0.48	1.89	0.69	1.31
18.00.00	0.62	2.16	0.69	1.50
18.15.00	1.48	1.65	0.69	1.14
18.30.00	2.20	1.47	0.69	1.02
18.45.00	1.98	1.27	0.69	0.88
19.00.00	1.88	1.15		1.15
⋮	⋮	⋮		⋮
23.45.00	3.84	1.55		1.55

It is also possible to set a condition which triggers the use of the adjustment, so that small differences are not adjusted since they can then be counterproductive. For example, PJM

have a requirement of the AM having to be greater than 1.05 and lower than 0.95 (a $\pm$ 5% difference) for the adjustment to be applied. Note also that various papers and methods use different windows to compute the AM. For example according to Grimm (2008), the NYISO use data from $t-3$ and $t-4$ to obtain their AM.

In our application we will be using the PJM multiplicative method[35] when making adjustments to our used baseline calculation methods, and since it has a requirement of the difference having to be over 5%, not all methods necessarily get adjusted. Mohajeryami et al. (2016) note reports that the choice of multiplicative or additive adjustment does not change the outcome substantially. We will, therefore, use the multiplicative adjustment, as the additive carries a caveat of the possibility of gaming by deliberately increasing load just before the curtailment period to boost the baseline (Xenergy, 2002). This makes the multiplicative method a bit better with combating the issue of gaming the system[36] (Wi et al., 2009) and moreover, ČEPS will also be using this method (ČEPS, 2022) making the decision all the more easier. In terms of the look-back window for AM calculation, we will use the values from two and three time slots prior to the DR event start.

5.5 Used Performance Metrics

To evaluate the precision of each method, generally some sort of performance metric is used. The most common performance metric when estimating the performance of a CBL calculation method is to calculate the mean absolute error (MAE). It is defined in Mohajeryami et al. (2016) as follows:

$$MAE = \frac{\sum_{i \in C} \sum_{d \in D} \sum_{t \in T} |b_i(d,t) - l_i(d,t)|}{|C| \cdot |D| \cdot |T|} \tag{14}$$

where C is the set of all customers, D is the set of all days and T is the set of all time slots in a day d. The lower the value for MAE is, the higher the accuracy for the method. In their codex, ČEPS states of the mean average percentage error (MAPE) when evaluating the accuracy. It is described in the ČEPS codex, as follows (eq. 15):

[35]Except for hp_4 in the low XofY and mid XofY methods, there we use an additive method due to a data anomaly. The anomaly being, that for that heat pump we have zero consumption for a large portion of the acceptable days vector, and thus these methods can return a zero baseline for the time slots used to calculate the AM. Thus, technically dividing a number by zero. The R software treats this value as infinity, which is why in these periods we use the additive adjustment instead of multiplying by infinity.

[36]Discussed more in section 7.

$$MAPE[\%] = \frac{1}{n} \cdot \sum_{1}^{n} \left| \frac{l_i(d,t) - b_i(d,t)}{l_i(d,t)} \right| \cdot 100 \tag{15}$$

where $l_i(d,t)$ represents actual consumption and $b_i(d,t)$ represents the baseline value at time t. Multiplying by 100 is done to obtain a percentage, and n stands for the number of fitted units.

Since ČEPS is working on an already aggregated level, it is possible to use the MAPE. This is because you most likely won't have actual or baseline loads being zero for any time slot. When doing it on individual heat pump level, you can have loads being zero for a time slot, which results in you dividing with or by zero, creating errors. This means that we will be using the MAE for evaluation, even though the ČEPS codex is otherwise used as a benchmark.

Additionally, we will use a method to calculate the bias of our CBL calculation methods. Mohajeryami et al. (2016) shows a changed formula of MAE that allow for non-absolute values for calculating the bias of the baseline method. The formula is changed to be the following:

$$Bias = \frac{\sum_{i \in C} \sum_{d \in D} \sum_{t \in T} (b_i(d,t) - l_i(d,t))}{|C| \cdot |D| \cdot |T|} \tag{16}$$

If the bias is positive the baseline method overestimates the consumers' actual load whereas negative values indicate the method is underestimating the consumers' actual load.

Both performance metrics will be used on non-DR days as well as DR event days (for event periods). Here we deviate slightly from the cited papers, as they don't generally test the non-DR days. However, as stated later as well, we feel that one can see the true performance of the method only if we compute the MAE and bias for periods in which there are no DR events, since we can expect the baseline and actual load to be different in event periods by definition.

5.6 Quantifying a DR event

When quantifying a DR event one must always bare in mind the context in which this is done. For example, if we look to industrial demand response, the event are often fixed in time and, event though this is not the end game, currently contracted. On the other hand, residential DR, by nature, is and has to be handled as dynamic, hence the number of events are not set and scheduled up front as in the industrial setting. The whole process of DR quantification can be summarised in three steps:

1. The DR signal is sent from the company facilitating the DR to the customer.
2. Each customer decides whether she would like to respond to the signal or not.
3. Using customer's smart meter data, the company reads the actual load and calculates the customer's incentive.

Moreover, as we state in section 4.2, Wijaya et al. (2014) define two types of DR signals, those being:

1. The signal information contains the start and end times of the event itself, and also the amount of kWh which are to be reduced.
2. The signal information contains the the start and end times of the event itself but lets the customer decide how much he is willing to reduce.

In this paper out signals are defined as the latter, so the individual customers can decide whether or not they let their HP decrease it's consumption. As Wiyaya et al. also state, this allows one to study how baselines and incentive allocation influence customers' decisions to reduce their consumption. However, after consulting the representatives supervising the thesis from ČEZ's viewpoint, this assumes, resp. works better, when one's DR scheme and the signal timing are based on these real time prices. Since our DR scheme is not, as it is based on one of the "typical daily diagrams" (TDD)[37] for residential heating there would most probably be a misalignment in the timing of the requests and thus our results would be bias, respectively influenced in a way we would not be able to fully interpretable and couldn't clearly state the causality of events. We therefore abstract from multiplying our kWh quantification of the DR scheme by the given day prices and

[37]Specifically the TDD 7 as shown here: `https://www.ote-cr.cz/en/statistics/electricity-load-profiles/normalized-lp?set_language=en&date=2022-05-14`.

restrain ourselves to reporting the results in kWh only. Still, as we discuss certain implications stemming from the quantification as described in Wijaya et al. (2014), we choose to briefly present their framework below. We present our simplified approach below.

5.6.1 Demand Response Quantification in Theory

Wijaya et al. (2014) propose the existence of three different loads during a DR event χ. An actual load $L(\chi)$, an estimated baseline $B(\chi)$ and a theoretical true baseline $B^*(\chi)$, which are the sums for all individual $l_i(\chi)$, $b_i(\chi)$ and $b_1^*(\chi)$. The "true" baseline is obviously unknown for the company, and can theoretically be known only by the customer, as they know what they would consume. As Wijaya et al. (2014) propose, the cost function for the company could be defined as a monotonically increasing and strictly convex function. For renewable resources this does not necessarily hold true, though it is noted that in the case for renewable resources, more expensive reserve generators may have to be activated to meet the high demand at a certain time. As an example that would satisfy these requirements, the following quadratic cost function is presented:

$$c(L) = a_1 L^2 + a_2 L + a_3 \tag{17}$$

where $c(L)$ is the total cost of meeting demand L and a_1 a_2 and a_3 are constants. Since the theoretical true baseline is unknown for the company, the perceived savings for the company for a DR event χ can be estimated by the following:

$$c(B(\chi) - c(L(\chi) \tag{18}$$

where the perceived saving is the difference between the cost of producing the estimated baseline and the cost of producing the actual baseline, at the duration of the DR event.

A customer's profit function is dependent on the reward a company would be willing to payout for participating in decreasing the load for the DR event period. The coefficient $\alpha \in [0, 1]$ in eq. 19 is defined as the proportion of savings the company is willing to pay out as incentives. The received individual profit for one customer can then be defined as

$$rp_i = \alpha \cdot (\frac{b_i(\chi)}{B(\chi)} c(B(\chi)) - \frac{l_i(\chi)}{L(\chi)} cL(\chi)). \tag{19}$$

This holds only if the actual individual load is lower than the estimated individual baseline. Otherwise the payout is 0. As an additional thing to note, since the customer could theoretically know their true baseline, their true individual profit could be estimated as:

$$tp_i = \alpha \cdot (\frac{b_i^*(\chi)}{B(\chi)} \cdot c(B(\chi)) - \frac{l_i(\chi)}{L(\chi)} \cdot cL(\chi)) \tag{20}$$

only if the actual individual load is lower than the true individual baseline. Otherwise the payout is 0. A customer's additional profit can then be shown as:

$$rp_i(\chi) - tp_i(\chi) \tag{21}$$

where if the value is positive, the customer i gets a higher reward than they would deserve.

The company's true profit cannot be calculated in reality, since it would require knowledge of the true baseline. Therefore we define the company's profit function as follows:

$$\frac{c(B(\chi)) - c(L(\chi)) - \sum_{i \in C}(rp_i)}{c(B(\chi))} \tag{22}$$

This means that a company's savings have to be calculated by using the estimated baseline, which makes it differ depending on which method is being used. This is why a company would want to use a method that produces the lowest MAE.

Since we don't have the data on prices, nor the data on potential reward paid out to customers, we will quantify the company's gain by calculating the difference of actual aggregated load compared to the estimated baseline by our most accurate method (and our benchmark method mid 4 of 6) for all DR events during Monday through Thursday for week 10, and note in the 7 that there are some possible additional prices and costs to take into account when calculating the whole profit. Additionally, we will extend the calculation to include the rebound, for a more accurate representation of actual load change owing to the DR scheme.

5.6.2 Our Quantification Approach

Our quantification period must be in alignment with the baseline estimation days, thus we quantify the DR for days between the 7^{th} and 10^{th} of March. We therefore have a

somewhat "representative" week that we can present results for. The theoretical base is explained in the previous section, however, it should be noted that due to several limitations, which are described in the following text and discussed in depth in section 7, we cannot provide a monetary representation of the DR scheme. Still, we believe that this doesn't greatly affect the implications of this paper, as the main focus here is the methods themselves along with their impact on quantification, and not the underlying DR scheme.

The main limitations, resp. reasons for not quantifying the respective customer and company profit and cost functions are the following. Firstly, we do not have the cost data to be able to truly quantify the cost functions, and even if we did, we would still run into the issue linked to dynamic prices being applied on a DR scheme with a "fixed" nature, resp. one which doesn't use real time prices for DR signals. This would cause the mentioned misalignment and bias. The cost functions could only be quantified in terms of parameters but since this papers main focus lies elsewhere, this would add value only if we could actually produce monetary results. Hence we abstract from this and focus on the results of the methods in terms of kWh.

Using a slightly simplified and modified equation 18, where we drop the cost function term, thus having only

$$B(\chi) - L(\chi), \tag{23}$$

we get estimation of kWh savings (or excessive use) for each the seven main method's baseline, including each methods adjusted baseline, thus giving us fourteen results. In equation 23, $B(\chi)$ is obtained for each of the four estimated workdays. The two terms in the equation can be defined as follows:

$$B(\chi)_d = \sum_{t=\lambda start}^{\lambda end} \frac{\sum_{i=1}^{C} b_{itd}}{C}$$

$$L(\chi)_d = \sum_{t=\lambda start}^{\lambda end} \frac{\sum_{i=1}^{C} l_{itd}}{C}$$

where $\lambda start$ and λend are the start and end times of the DR event, $t \in T$ are the days time slots, $i \in C$ and the individual heat pumps (customers), $d \in D_{DR}$ are here not all the acceptable days, but instead our four estimated DR event days. One can then sum

the individual day results to obtain the full weeks DR quantification.

Moreover, as we believe than one should never omit the quantification of the rebound from any form of DR assessment, we also repeat this for rebound duration hours. Prior to rebound quantification, it is crucial to set the conditions defining the rebound effect. As the main focus of this paper is not the DR scheme itself, we choose a simple rebound identification mechanism. In short, if the rebound starts within 3 time slots from the end of the DR event it is included and we do not consider deviations after this time limit to be rebound effects. The rebound ends when the actual load line and the estimated baseline intersect again, or with an upcoming event which can happen in our data set due to the presence of multiple occurring events during the day. Since the effects of including the rebound are done for demonstrational purposes, the extension of adjustment and quantification time is done based on the average rebound length. The chosen times are clearly visible in the plots in section 6.3.

6 Results

In the following section results are presented. These are the performance metrics, method comparison, DR quantification with and without considering the rebound effect after the event. In order to be able to offer at least a snapshot of the DSM scheme evaluation, we choose to estimate the baseline for all heat pumps and take both their average load and average estimated baseline to show the actual available flexibility representative. The week we chose to test our methods are the workdays of week 10 in 2022. This is for a few reasons. The 9th of March, Wednesday in week 10 has the highest summed load (consumption) in the data set. Having the week in March also provides us with enough acceptable days to do estimations with. If one were to go further forward from this week, the only acceptable days added in our vector would be Fridays, giving our estimations a bigger bias towards Friday load curve data.

6.1 Method comparison

For finding the method with the best performance metric, we choose to calculate and compare the MAE and Bias for 5 non-DR days, the 4^{th}, 11^{th}, 18^{th}, and the 25^{th} of March. The reason for this is that including DR events skewes these results, because a good method could seem to have a high bias and MAE, because it correctly differentiates the baseline from the actual load during DR events. This means that doing these metrics on non-DR days is closer to comparing the estimated baselines to true baselines. The issue here is that methods with adjustments cannot be measured, since the adjustments are made based on DR events. One can therefore assume that most methods, especially the simple averaging methods, perform better with adjustments included.

In figure 10 we can see how the other methods compare to our benchmark of Mid 4 of 6. Surprisingly, the benchmark method seems to have the highest MAE of all. But, the figure also shows some things that are to be expected. The High 5 of 10 has a positive bias, meaning that the method usually overestimates the baseline compared to actual load, whereas the Low 5 of 10 underestimates it. The Mid 4 of 6 then having having a very small bias. These methods have the three highest mean absolute errors, which means that they on average estimate the model incorrectly, where the estimate evens out for the Mid 4 of 6 in the long run due to the low bias, but not for the High and Low 5 of 10. When looking at bias the linear regression, EMA and the Exponential

Smoothing model shine through the most. This means that even though these methods do on average have an error per time slot, when estimating a whole day, it measures the complete usage very closely. This is a result that you could get for instance if the estimate tracks the actual perfectly but the actual load curve includes a lot of spikes, whereas the estimated baseline is rather steady. The exponential moving average and the exponential smoothing method are both very similar, with the exponential moving average having a slightly smaller bias, but the exponential smoothing method having a slightly smaller MAE. The Regressions have the lowest MAE, with the Linear Regression obtaining the lowest by far. The Linear Regression also clearly has a lower bias than the Quadratic Regression, rendering the Linear Regression the superior method based on our performance metrics. The reason for the Linear Regression performing so well compared to the quadratic regression, could be that for week 10, the temperatures did not go over 10.4 °C. If there is to be a parabolic relationship between temperature and electricity consumption one could assume that the parabolic curve starts growing monotonically in days with generally warmer weather conditions. This assumes the installed heat pump is capable of both heating and cooling, which is the case for our sample. These metrics were also computed for a non-DR weekend, specifically the 26^{th} and 27^{th} of February (see appendix 8). Interestingly, for weekends, the lowest MAE is produced by the Exponential Smoothing. However, the Linear Regression still has the lowest bias, and with it also having the second lowest MAE, one could still conclude that it is the superior method. Moreover, since it only uses a very limited amount of days here (see table 18 in appendix 8) its performance will most probably grow with the amount of usable observations.

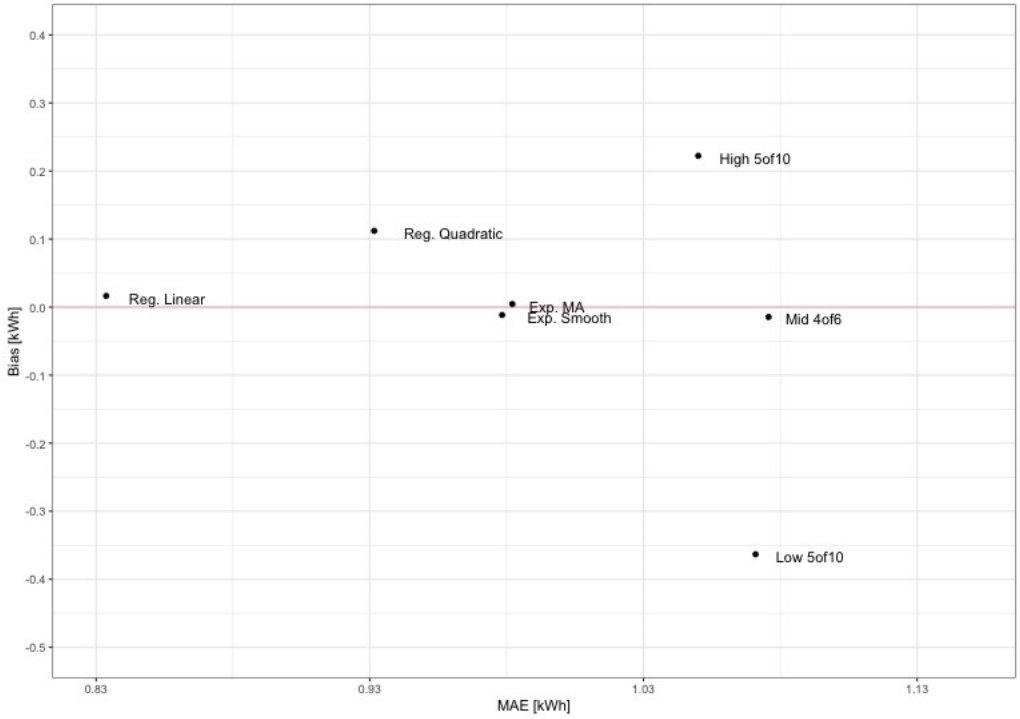


Figure 10: Performance Metrics for non-DR Days

Table 11: Performance Metrics non-DR Days

Method	MAE (kWh)	Bias (kWh)
Mid 4of6	1.0757	-0.0145
High 5of10	1.0500	0.2224
Low 5of10	1.0710	-0.3633
Exp. Moving Average	0.9821	0.0045
Exp. Smoothing	0.9783	-0.0116
Linear Regression	0.8337	0.0164

Looking at the MAE and Bias of all of the methods in figure 11 (resp. table 12), during DR event periods in week 10 one can see indications of potential DR scheme quantification results. For example, if we were to combine the results of our performance metrics on both non-DR and DR periods, we could conclude the following. If we assume the linear regression method to be the most accurate, based on figure 10 (resp. table 11, then after observing it's bias during the DR periods (figure 11 or table 12) we could say it indeed lowers overall consumption for measured days. Looking to the benchmark method (Mid 4 of 6) we see that for DR days it would report an overall increase (negative bias). Interestingly enough when adding adjustments, these two results become inverted. This is touched upon mainly in section 6.2 where the results of the quantification are presented.

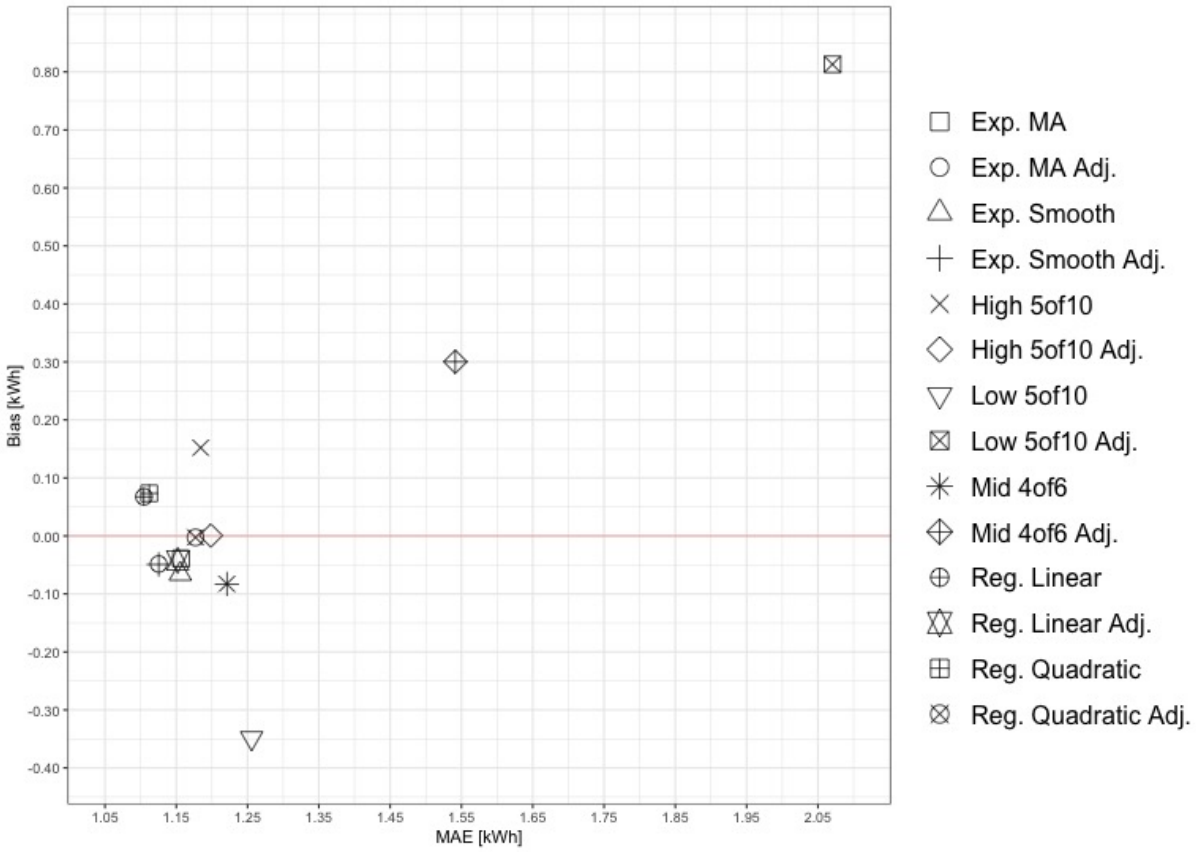


Figure 11: Performance Metrics for DR Event Periods – Including Adjustments to Methods

Table 12: Performance Metrics for DR Event Durations

Method	MAE (kWh)	Bias (kWh])
Mid 4of6	1.2210	-0.0828
Mid 4of6 Adj.	1.5414	0.3003
High 5of10	1.1842	0.1520
High 5of10 Adj.	1.1985	0.0006
Low 5of10	1.2557	-0.3486
Low 5of10 Adj.	2.0704	0.8131
Exp. Moving Average	1.1568	-0.0389
Exp. Moving Average Adj.	1.1254	-0.0485
Exp. Smoothing	1.1556	-0.0655
Exp. Smoothing Adj.	1.1253	-0.0492
Linear Regression	1.1047	0.0676
Linear Regression Adj.	1.1521	-0.0423
Quadratic Regression	1.1121	0.0734
Quadratic Regression Adj.	1.1770	-0.0026

Moving forward, based on the performance metrics in figure 11, we will show the process of quantifying DR with the Linear Regression and our benchmark method Mid 4 of 6, along with their respective adjustments. In figures 13 and 14 we find the average baseline estimation along with adjustments for the estimated week for the methods Mid 4 of 6 and Linear regression respectively. We can see that the adjustments for the Mid 4 of 6 method do not always work that well. This is because the adjustments found in these figures are the average adjustments from all heat pumps. If the estimated baseline vastly differs from the actual load, the adjustment can become huge for some individual heat pumps due to the properties of the multiplication method. For the Mid 4 of 6 method this can happen frequently on individual heat pump level. For demonstrational purposes, we find the estimated baselines of the two different methods along with their adjustments for the Tuesday of the estimation week for a randomly selected heat pump in figure 12.

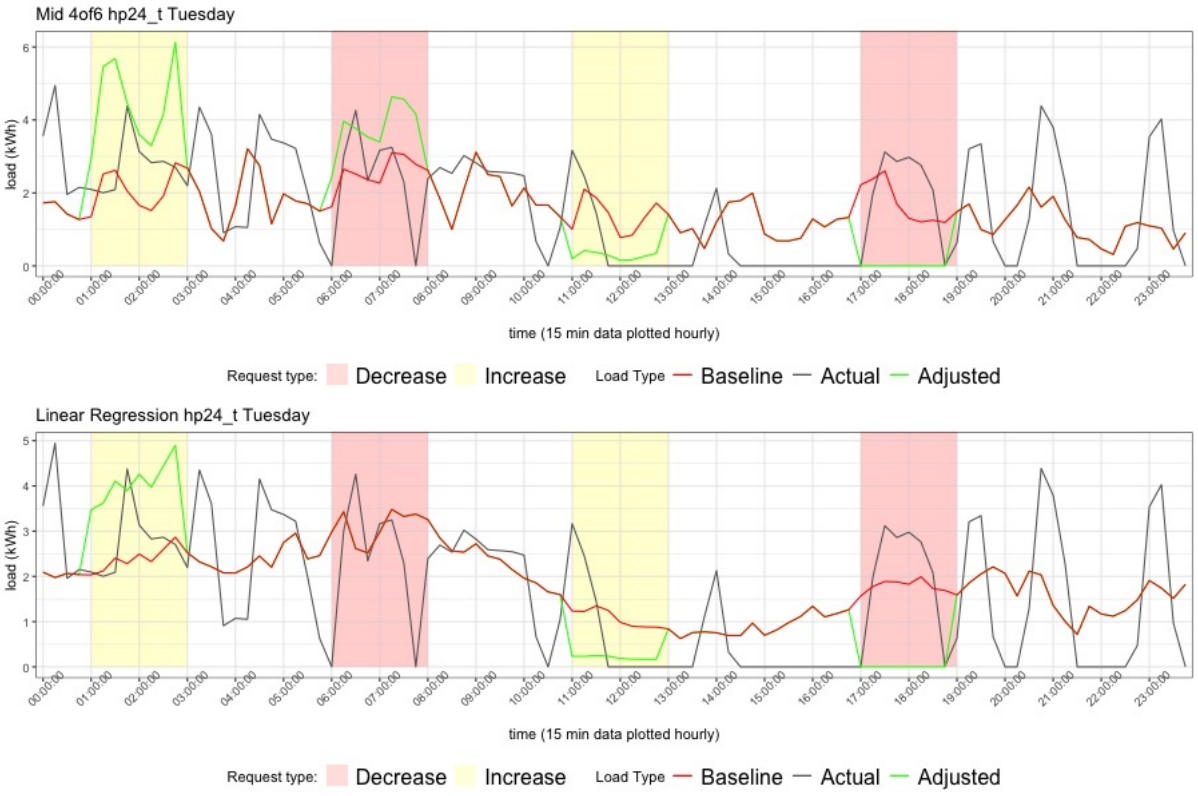


Figure 12: Single Heat Pump Example – Actual load plotted against baseline and adjusted baseline for the DR event and rebound.

Here we see that due to the properties of the AM, the value of the AM becomes quite large when there has been a spike in consumption two to three time slots before the DR event. This makes the adjustments quite substantial for some events. In turn, the averaged adjustments from each heat pump look bigger compared to just making the adjustments on an already aggregate level. The next section will go over the quantification of DR.

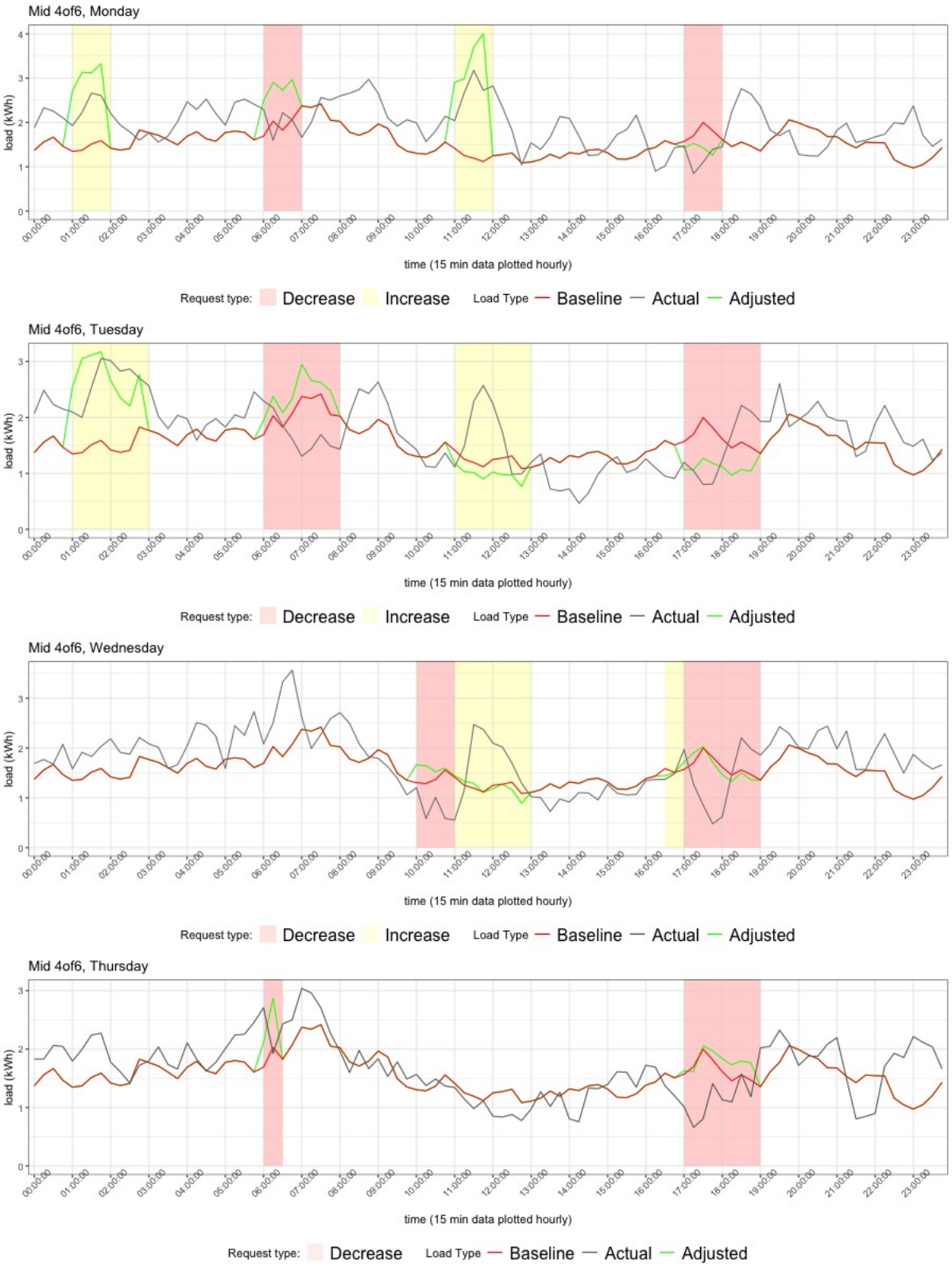


Figure 13: Mid 4of6 – Actual load plotted against baseline and adjusted baseline for DR event only.

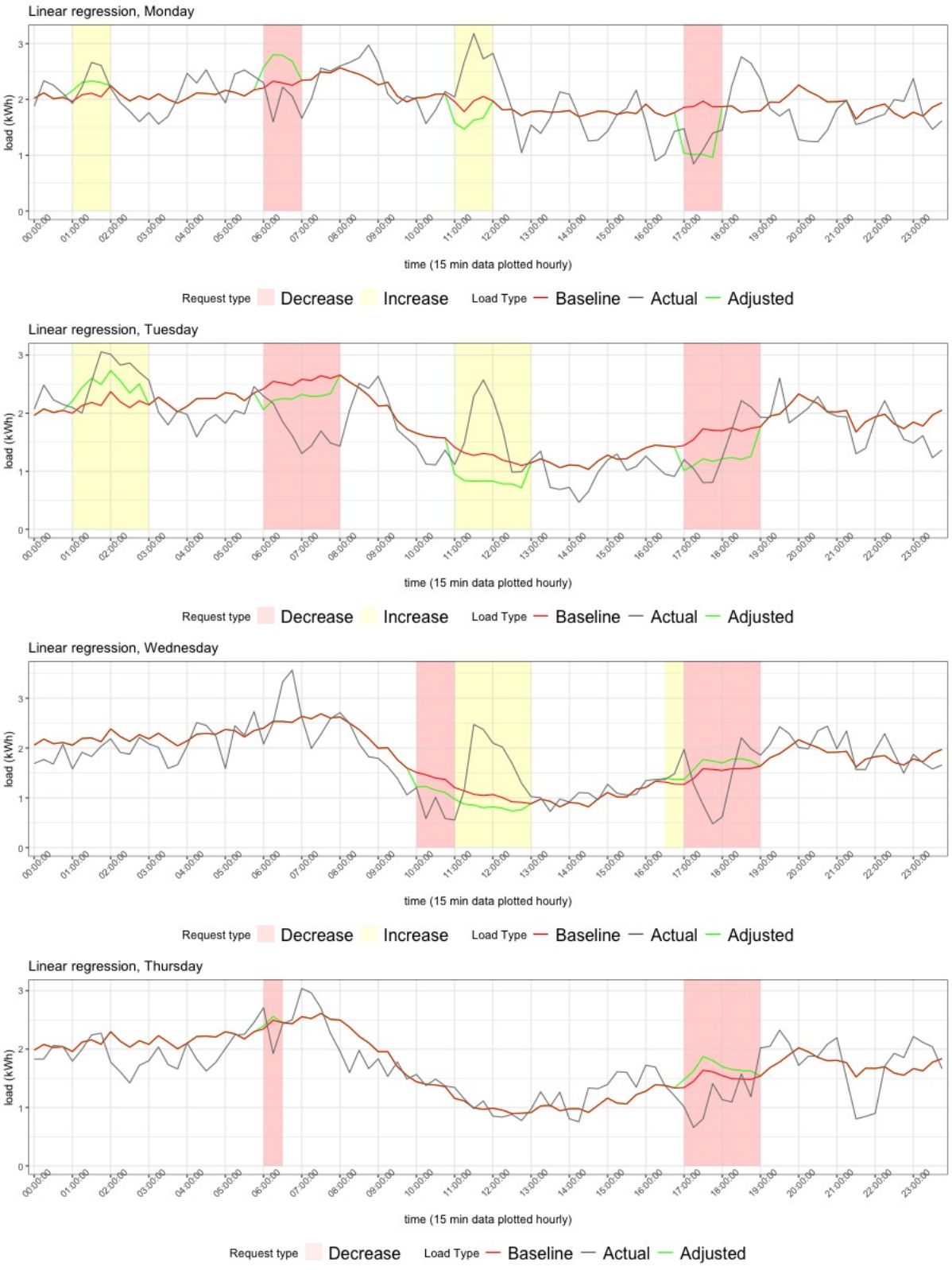


Figure 14: Linear regression – Actual load plotted against baseline and adjusted baseline for DR event only.

6.2 Quantification

Table 13 shows how the quantification of the DR scheme would look like in total kWh for week ten[38] as the total sum from each DR event. As one would expect, the increase requests, depicted with an (I), show a total increase in consumption, where as the decrease requests, depicted with a (D), show a decrease in consumption. With more decrease requests than increase requests (see table 3), one could expect the total sum to be positive (indicating kWh decreased) for the whole DR scheme. As mentioned above, the adjustments can cause the value sign to be the opposite of what it was for the method before the adjustment, indicating possible overcompensation. Therefore, to tackle these over-compensations, it might be more prudent to aim for a method which generates small bias to start with. This again points to the regression methods, as in their case, the conditional adjustment often doesn't even occur, and if so, it's adjustment multiplicative tends to be lower, meaning a lesser chance of possible overcompensation. Based on the literature review, it is more common to quantify only the duration of the DR event. Doing quantification this way doesn't take into account the possible counter reaction to the DR event, the rebound. While the consumption after the DR event isn't strictly decided by the DR event, it is most definitely affected by it, which is why we incorporate the rebound consumption when quantifying DR in the next section.

Table 13: Quantification for Benchmark and Best Method – Only DR Event Duration

Method & Request Type	Mon	Tue	Wed	Thu	Method Sum (kWh)
Mid 4of6 (D)	9.45	27.42	24.54	18.65	80.05
Mid 4of6 (I)	-50.80	-70.27	-19.33	0.00	-140.40
Sum (kWh)	-5.76	-72.14	-15.73	96.12	-60.35
Mid 4of6 Adj. (D)	20.67	17.79	28.69	32.34	99.49
Mid 4of6 Adj. (I)	32.25	-26.47	-21.29	0.00	-15.50
Sum (kWh)	-5.76	-72.14	-15.73	96.12	83.98
Linear Regression (D)	20.09	47.63	19.90	18.57	106.19
Linear Regression (I)	-22.38	-39.40	-30.54	0.00	-92.33
Sum (kWh)	-5.76	-72.14	-15.73	96.12	13.86
Linear Regression Adj. (D)	10.30	13.99	21.55	26.57	72.40
Linear Regression Adj. (I)	-25.34	-42.82	-39.24	0.00	-107.40
Sum (kWh)	-5.76	-72.14	-15.73	96.12	-35.00

[38]Note that we quantify only the workdays as weekends are estimated used modified versions of the same methods for reasons described in section 5 and thus we abstract from their estimation.

6.3 Rebound Effect

When taking rebounds into account, both the adjustments and time of event quantification have been extended according to the criteria in section 5.6. Once including the rebound, we can see that the kWh savings resulting from the DR scheme are in total positive only for the unadjusted linear regression method. And even this methods final sum is lower when compared to when include only the duration of the DR event. All other methods would indicate that the DR scheme causes more consumption in kWh, if one takes the rebounds into account. Whether or not complete savings in kWh is strictly good or not is further discussed in section 7. Nevertheless, our results indicate that taking rebound into account can indeed change the out-turn when quantifying the value of DR and thus should not be omitted from DR quantification. In figures 15 and 16 above we see the changed adjustments based on the criteria mentioned above.

Table 14: Quantification for Benchmark and Best Method – Including Rebound

Method & Request Type	Mon	Tue	Wed	Thu	Method Sum (kW)
Mid 4of6 (D)	-53.80	9.91	-113.00	-59.45	-216.34
Mid 4of6 (I)	-50.80	-51.88	-113.00	0.00	-215.67
Sum (kW)	-104.59	-41.96	-226.00	-59.45	-432.01
Mid 4of6 Adj. (D)	12.88	6.25	-114.89	-11.99	-107.76
Mid 4of6 Adj. (I)	32.25	-23.04	-114.89	0.00	-105.68
Sum (kW)	45.13	-16.79	-229.78	-11.99	-213.44
Linear Regression (D)	4.11	48.95	12.79	5.58	71.43
Linear Regression (I)	-12.31	-12.50	-30.34	0.00	-55.15
Sum (kW)	-8.20	36.45	-17.55	5.58	16.29
Linear Regression Adj. (D)	-0.77	-2.51	18.71	20.30	35.72
Linear Regression Adj. (I)	-16.59	-14.57	-41.94	0.00	-73.10
Sum (kW)	-17.36	-17.08	-23.23	20.30	-37.37

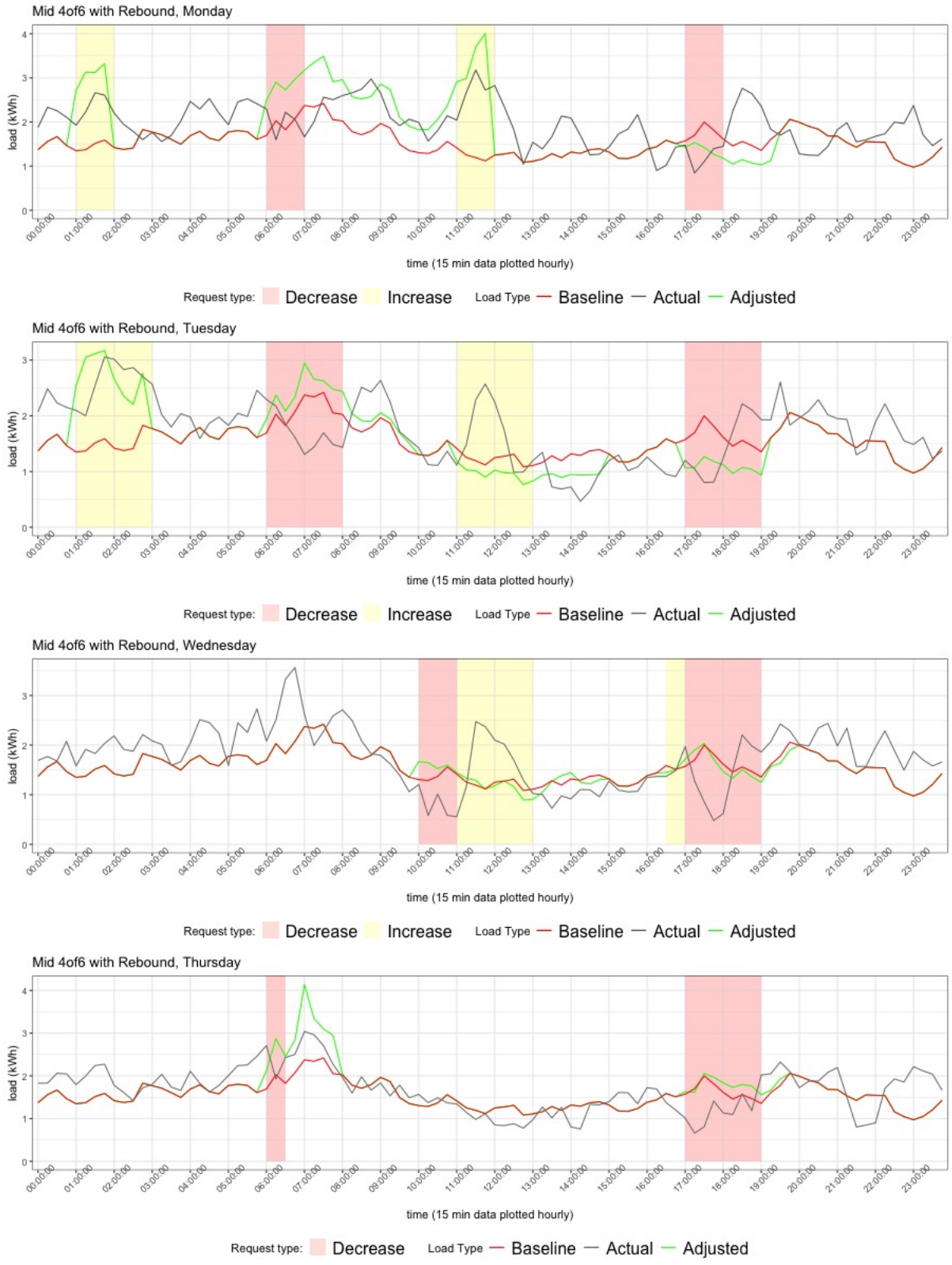


Figure 15: Mid 4of6 – Actual load plotted against baseline and adjusted baseline, including the rebound effect.

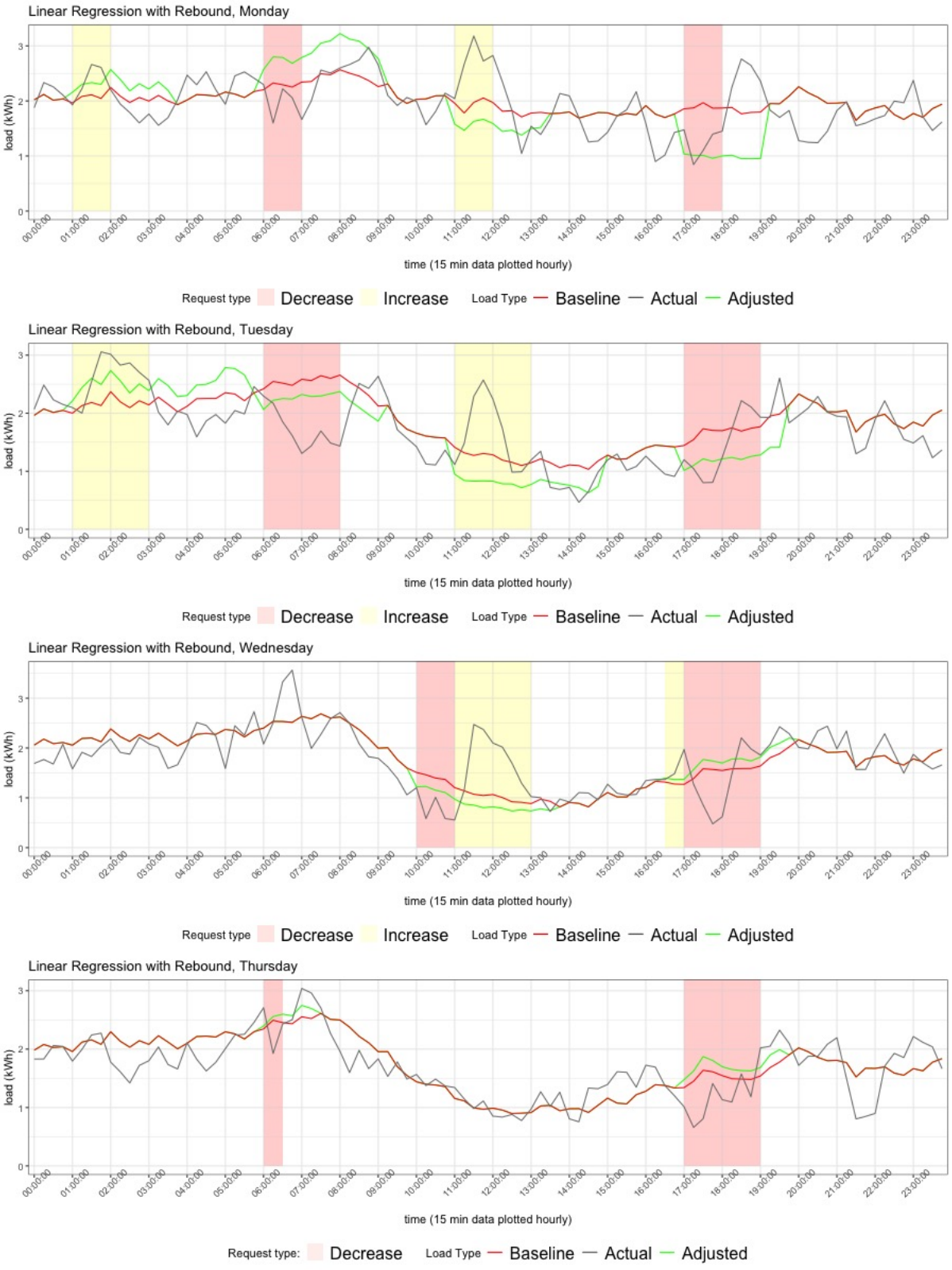


Figure 16: Linear regression – Actual load plotted against baseline and adjusted baseline, including the rebound effect.

7 Discussion

In this section we present discussion points that have not been fully explained in the results section, as well as expand on some limitations and further research suggestions. Also, before commenting further we would like to point out one last time the framework of this paper. Our aim is not to quantify and evaluate the underlying DR scheme in full, as our principle topic are the baseline methods and their workings themselves. That being said, we attempt to show how the used methods would perform if we applied them in a simplified quantification methodology to show a glimpse of how far the implications of baseline calculations might go and why we are actually doing them. The answers to why are always dependent on the participant who's viewpoint we wish to study (see section 2). In this paper we are interested mainly in the balancing market, which ČEZ a.s., being the company providing the data and framework, will be interacting in. Thus, the deviations of our methods are of great concern as a well defined and tweaked baseline method has the potential to become a powerful hedging tool. For illustration, if we look at the website of the Czech TSO ČEPS[39], we can find the prices of deviations in CZK/MWh. Taking the period of the last six months (from early October to late March) the highest hourly price which occurred was 26,547 CZK per MWh. Although this value occurred during the volatile times in last falls energy turmoil, we use such an extreme value to illustrate the magnitude of potential risk that DSOs and retailers of energy can find themselves facing. Hopefully, these extreme prices won't be the standard of the upcoming years, however, the volatility in energy production that will inevitably accompany the transition away from traditional sources of energy will be a great challenge. Thus, we advocate the idea of giving participants of this system tools by which they can ensure some kind of price stability, and thus be able to at least mitigate and overcome the short term negative externalities of this transition if most important. Demand response (resp. DSM) and the overall flexibility of certain load is one of these tools, but without proper baseline estimation there is currently not a way how to make these tools usable. We would therefore like to stress the importance of focusing on this small but important piece of the greater puzzle that is DSM.

To continue with this section we discuss the reasons for our regressions working so well in our study. First off, one has to remember that the electricity usage measured in our data is for heat pumps specifically. Heat pump usage can be assumed to be highly correlated to the outside temperature, which our regressions are based on. What this means is

[39] https://www.ceps.cz/en/all-data#OdhadovanaCenaOdchylky

that weather based regressions aren't unquestionably the best for estimating baselines for complete household electricity usage, even though the regression methods prove to be the best for our data. Based on some of the poor adjustments created by the multiplicative method and the relatively good performance of the regressions, one could suggest that perhaps an averaging method that uses weather based regressions for adjustments could be a decent option for baseline estimations for complete household electricity usage. But, if one would want to keep a DR scheme separate for heat pump usage specifically, using a weather based regression method for estimating the baseline seems like a reasonable option based on our study.

As mentioned earlier, the linear regression seems to be superior to the quadratic regression. We would still recommend considering using a quadratic regression, if one is to make baseline estimations for heat pump usage for a whole year. This is due to the fact that for heat pumps that possess both heating and cooling capabilities, the relationship between outside temperature and consumption has signs of being parabolic. The reason for the linear regression performing so well in our study could be that the data is from winter to early spring, where the temperature for our DR event week[40] ranges from -4.6 °C to 10.4 °C, where the relationship is still at the former half of the parabola, which is downward-sloping. When having the regression run for a whole year, the temperatures get higher which would move us into the upward-sloping area of the parabola, rendering it difficult for a linear regression to be exact. The is a risk of a linear fitted line being a flat horizontal line if a whole years data is included, whereas for the short winter/early-spring period the linear regression doesn't necessarily have this same issue. If one would want to still use a linear regression, one would most likely have to split up the regression period into seasons during the year.

One could also make the argument that due to the effects of the pandemic, people spend more time at home, which gives room to the question of whether or not differentiating between workdays and weekends/holidays is necessary. While it is true that people's daily routines have changed due to the pandemic, we still believe that people are more likely to be out of their home during weekends/holidays and usage of electricity in total still varies between people's work and leisure time, which is why we still chose to differentiate between the two.

[40] Otherwise, for the whole sample the temperature range is wider, see section 4.

As for limitations of this paper, due to limited data, some other more experimental methods were not able to be done. This includes a method of Synchronous Pattern Matching[41] proposed by Wang et al. (2018). It is a method where a group of control customers are grouped into clusters by the method of k-means. The cluster is then matched to a DR participant based on the similarity of the actual load for the time before and after the DR event. This would have been an interesting method to include, but with our data only having 5 control heat pumps compared to 22 DR participant heat pumps, it was simply unfeasible to do the method. The reason for not obtaining a larger control sample here are certain technical issues. Additionally, even though this method is interesting and possibly results in low errors, it would demand the majority of customers not participating in a DR program, which one could think that in the long run is counterproductive to the goal of DSM. If one was aiming for a fully flexible household demand, one wouldn't necessarily want a rather large portion of households not participating in the flexibility program.

For further research, Weng et al. (2018) mention that for residential users it might be better to use probabilistic methods for baseline estimation since residential consumer usage is more more volatile and irregular compared to commercial customers, thus adding stochastic terms could represent the baseline more accurately. The problem is that implementing a method like this requires an enormous amount of "fine grained data" (Sun et al., 2019a) in addition to more computational power Weng et al. (2018), which is something we lack in our case.

Another limitation of SPM and probabilistic methods are that by using those methods an operator could run into the problem of getting people involved in DSM, since the customers have to understand and trust the quantification that the operator does. It is important that an operator uses transparent and understandable methods (ENERNOC, 2011), where the simpler averaging methods shine through. An incredibly sophisticated and complicated method could theoretically produce more precise results, but in the end employing that method could cause people not to want to participate, rendering it not as profitable. One has to keep in mind that words like Synchronous Pattern Matching, Probabilistic and Stochastic can possibly evoke distrust and averseness in an ordinary customer. Especially in current day Czechia where people recently witnessed one of the larger retailers, Bohemia Energy, going out of business[42], due to the combination of the

[41] a more detailed summary of SPM can be found in the appendix

[42] See for example here `https://www.reuters.com/world/europe/czech-firm-bohemia-energy-shuts-down-citing-surging-power-prices-2021-10-13/`

late price volatility and lack of hedging in combination fairly risky portfolio management. Electricity bills have therefore become an even hotter topic. The effects of method sophistication on willingness to participate in DR might be and interesting topic for behavioural economists. Additionally, computational time plays a big role in method selection (Weng et al., 2018) if one wishes to do these estimations in real-time. Due to the combination of gaining people's trust with understandable methods, in addition to the limitations of computational time and data, we assume that ČEPS may have chosen the averaging method of Mid 4 of 6 for these reasons.

When it comes to using adjustments, the problem of gaming the system was mentioned earlier. When using additive and multiplicative adjustments, a customer can increase their consumption for a short moment before the DR event, showing then an exaggerated change in consumption during the DR event, if the adjustment was made on wrongful basis. The additive method is even more vulnerable to this, due to the simplicity of the additive adjustment, though the multiplicative method isn't immune to this either. Making an adjustment based on a regression with a weather component could then be better at combating this issue. The gaming issue stems from giving customers additional monetary incentive to participate in a DR scheme. Here a company falls into a dilemma of wanting to get more people to participate via monetary compensation, while running the risk of customers gaming the system. In the quantification methodology presented in section 6.2, where we present the compensation scheme proposed by Wijaya et al. (2014) one can see how a customer could essentially manipulate their received profit. Since the incentive paid out to each individual customer is based on the estimated baseline, a customer can deliberately increase their consumption for a short period before the DR event, substantially changing the estimated baseline from the theoretical true baseline. When the customer does this, not only are excessive amounts of energy used before the event, but the incentive that the customer gets paid becomes inflated. This is why one may consider using adjustments based on weather data instead of consumption prior to an event. An additional issue with adjustments based on the consumption of a few time slots prior to the event are that the consumption a few time slots prior could vastly differ from the consumption right at the event, causing in an incorrect adjustment. Choosing the correct amount of time slots to take into account when making an adjustment like this can therefore vary depending on the situation. Even so, looking at the shortcomings that simple averaging methods can have when comparing their performance metrics with the more sophisticated methods, adjustments do seem to be an important piece in making

simple averaging methods more precise.

When it comes to quantification of a DR scheme, there are a few things to take into account. When choosing the period for which you want to value a DR scheme, you can go with different options depending on your preferences. You could make a quantification for each week separately, you could do it for each workday week and each weekend separately, or you could do it for each day separately. We choose to isolate the workday week, as that is how the methodology differentiates between them. When valuating the DR scheme as a whole, you have to then sum together all of your quantified periods to see the value of your chosen interval. This interval will most probably be determined by the chosen settlement frequency DR participants, which will probably differ for every party involved. For residential customers it could make sense to link this with their monthly electricity bill, whereas for DSOs a daily or weekly basis could make more sense. The final choice should be based on the existing infrastructure in order to be as much aligned with current workings of the system as possible. When simplified, if we would strictly look at the equations 17 and 18 proposed by Wijaya et al. (2014) one could assume an actual load lower than the estimated baseline would result in more savings for the company. Additionally, in our case, one could assume that the participation in a DR scheme, which includes more decrease requests than increase requests, should result in a decrease of total kWh usage, so as to result in savings for the environment and the company. With the data available for our study, this is how one could simply quantify the value of a DR scheme. In reality, there are more things to take into consideration. Energy prices on the real world market are dynamic, which means that depending how someone chooses to time their consumption for a day, both a customer and a retailer can win or lose money. It is known that the energy production of renewable resources peak at certain times of the day. This means that producing electricity on different times of the day at different capacities, can affect that periods energy mix. For example, when there is a significant lack of supply for a short period, expensive and potentially environmentally unfriendly backup energy generators may have to be activated to fulfil the demand. Satisfying this short burst of demand can result in a lower total electricity consumption for a day, but the electricity used could be more expensive and more damaging to the environment. This means that when simply quantifying a DR scheme with the data at our hands, we cannot know for certain if the result of either saving in kWh or using more kWh is good or not. We would need data from the whole system to be able to make this statement. Since the consumption during a DR event is affected on purpose, one could assume that a total

increase in kWh consumption caused by a DR scheme could be more evened out throughout the day, making it cheaper and giving opportunity for the usage of renewable sources. Nonetheless, our results show that the accuracy of the method can substantially change the outcome of quantification (changing the value from positive to negative). Besides, the DR events weren't based on real-time electricity prices, but pre-defined by TDD tariffs (see section 5.6). This could result in a DR event proving to be unprofitable when taking real-time prices into account, even though the scheme made sense when looking at the tariffs. Either way, using anything other than real-time prices to make the quantification would be faulty. As an addition, satisfying a short burst of demand could be cheaper and result in a smaller kWh usage overall, but with the production being more damaging to the environment. Here a retailer falls into the dilemma of prioritising between the environment and short term profit. Though, these situations are something that could be handled by tariffs and regulation.

DR schemes can also cause short bursts of high demand. This is either an increase request, or a possible rebound from a decrease request. A short burst of demand caused by an increase request is most often controlled, and the retailer knows that the consumption can be met sufficiently. As for a rebound effect, that may not always be the case. A DR scheme can result in peak shaving for a DR event period, but the rebound after could still prove to be problematic. That is if a rebound caused by a decrease request has a higher peak than the consumption curve that would've happened without intervention. Since a possible rebound effect is directly caused by a DR event, we would recommend the inclusion of rebounds when quantifying the effects of DR. As our results show, when including the rebounds of the DR events, the kWh amounts gained or lost clearly change. Again, to surely estimate whether or not the DR scheme was valuable or not, one would both need to consider dynamic prices, the energy sources and the individual parties' viewpoints. Even so, taking the rebounds into consideration seems to change the outcome.

Lastly, based on the discussion above, as a proposal for further research we suggest finding reliable and exact methods for determining the exact durations and sizes of rebounds for more accurate estimation and quantification of DR.

8 Conclusion

In this thesis we compared different baseline estimation methods as well as quantified DR for a regional data set in Czechia. Finding an optimal method for baseline estimation, as well having the correct approach to quantifying DR is an important cornerstone of demand side management, which in turn is vital for balancing and optimisation of the energy market as a whole. Research on this data set had not been done before, as it is a part of a pilot project by the Czech energy conglomerate ČEZ a.s. Thus, this can be seen as the main contribution of our paper.

The methods chosen for baseline estimation were High-, and Low 5 of 10, EMA, Exponential Smoothing, Linear Regression, and Quadratic Regression along with the benchmark method of Mid 4 of 6. We also carry out adjustments for each of these methods, specifically the multiplicative adjustment. This means we effectively have 14 methods to compare. Our estimated workday week is week 10 in 2022. In this week, all of the alternative methods had lower MAE when compared to the benchmark method. The EMA, Exponential Smoothing and the Linear Regression all had very low biases, along with the benchmark method, the rest had visibly higher biases. In the end, the Linear Regression proved to be the most accurate method. Our results for quantification of DR show that depending on the accuracy of the model, the implications of the quantification can substantially differ. Moreover, we conclude that by including an estimate of the rebound when quantifying the DR the overall outcome can change, if compared to only quantifying the duration of the DR event itself. Lastly, we also conclude that in order to properly estimate the value of DR, one would need to find a working alignment between using the dynamic price data for the specific electricity market, data on which sources are penetrating the market at every given point in time, the workings of tariffs in place, and the actual DR scheme valuation itself. The final evaluation of a DR scheme very much depends on the chosen viewpoint and preferences or goals this viewpoint holds. We also offer several further research propositions, all of which however inevitably build on, or use, the methodology and approaches listed in this paper.

With the intrigue of having a regional specific sample also came a downside of the sample being small and for a relatively short time period. This also leaves room for future research as putting these methods to test alongside those used here in a wider data setting would give new insights and pave the way to finding the most reasonable baseline estimation method. The methods being, e.g. Synchronous Pattern Matching and various

probabilistic methods. The former requires a larger set of control observations than what was available to us at this given point in time, whereas the latter requires huge amount of fine grained data and computational power. Additionally, we would suggest finding optimal methods for estimating the rebound most accurately for further research.

Bibliography

Babar, M., Nguyen, P. H., Ćuk, V., Kamphuis, I., Bongaerts, M., and Hanzelka, Z. (2017). The evaluation of agile demand response: An applied methodology. *IEEE Transactions on Smart Grid*, 9(6):6118–6127.

Bampoulas, A., Saffari, M., Pallonetto, F., De Rosa, M., Mangina, E., and Finn, D. (2019). Quantification and characterization of energy flexibility in the residential building sector. In *The 16th International Building Performance Simulation Association (IBPSA 2019), Rome, Italy, 2-4 September 2019*. International Building Performance Association.

Bertoldi, P., Zancanella, P., Boza-Kiss, B., et al. (2016). Demand response status in eu member states. *EUR 27998 EN*, pages 1–140.

Bliek, F., van den Noort, A., Roossien, B., Kamphuis, R., de Wit, J., van der Velde, J., and Eijgelaar, M. (2010). Powermatching city, a living lab smart grid demonstration. In *2010 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT Europe)*, pages 1–8. IEEE.

Brusco, G., Burgio, A., Menniti, D., Pinnarelli, A., and Sorrentino, N. (2014). Energy management system for an energy district with demand response availability. *IEEE Transactions on Smart Grid*, 5(5):2385–2393.

Council of European Union (2019). Council regulation (EU) no 943/2019. `https://eur-lex.europa.eu/eli/reg/2019/943/oj`.

Devriese, J., Degrande, T., Mihaylov, M., Verbrugge, S., and Colle, D. (2019). Techno-economic analysis of residential thermal flexibility for demand side management. In *2019 CTTE-FITCE: Smart Cities & Information and Communication Technology (CTTE-FITCE)*, pages 1–6. IEEE.

ENERNOC (2011). "the demand response baseline" white paper. Available at: `https://library.cee1.org/sites/default/files/library/10774/CEE_EvalDRBaseline_2011.pdf`.

European Smart Grids Task Force (2019). Final report: Demand side flexibility – perceived barriers and proposed recommendations. `https://ec.europa.eu/energy/sites/ener/files/documents/eg3_final_report_demand_side_flexiblity_2019.04.15.pdf`.

Foteinaki, K., Li, R., Heller, A., and Rode, C. (2018). Heating system energy flexibility of low-energy residential buildings. *Energy and Buildings*, 180:95–108.

Grimm, C. (2008). Evaluating baselines for demand response programs. In *AEIC Load Research Workshop*, pages 1–31.

Haas, R. and Biermayr, P. (2000). The rebound e!ect for space heating Empirical evidence from Austria. *Energy Policy*, page 8.

Hatton, L., Charpentier, P., and Matzner-Lober, E. (2016). Statistical Estimation of the Residential Baseline. *IEEE Transactions on Power Systems*, 31(3):1752–1759.

Hedegaard, K. and Balyk, O. (2013). Energy system investment model incorporating heat pumps with thermal storage in buildings and buffer tanks. *Energy*, 63:356–365.

Hedegaard, R. E., Kristensen, M. H., Pedersen, T. H., Brun, A., and Petersen, S. (2019). Bottom-up modelling methodology for urban-scale analysis of residential space heating demand response. *Applied Energy*, 242:181–204.

Jazaeri, J., Alpcan, T., Gordon, R., Brandao, M., Hoban, T., and Seeling, C. (2016). Baseline methodologies for small scale residential demand response. In *2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)*, pages 747–752, Melbourne, Australia. IEEE.

Jin, M., Feng, W., Liu, P., Marnay, C., and Spanos, C. (2017). Mod-dr: Microgrid optimal dispatch with demand response. *Applied energy*, 187:758–776.

Kerscher, S. and Arboleya, P. (2022). The key role of aggregators in the energy transition under the latest european regulatory framework. *International Journal of Electrical Power & Energy Systems*, 134:107361.

Khabdullin, A., Khabdullina, Z., Khabdullina, G., Lauka, D., and Blumberga, D. (2017). Demand response analysis methodology in district heating system. *Energy Procedia*, 128:539–543.

Klaassen, E., van Gerwen, R., Frunt, J., and Slootweg, J. (2017). A methodology to assess demand response benefits from a system perspective: A Dutch case study. *Utilities Policy*, 44:25–37.

Kok, K. (2013). The powermatcher: Smart coordination for the smart electricity grid. *TNO, The Netherlands*, pages 241–250.

Koliou, E., Eid, C., Chaves-Ávila, J. P., and Hakvoort, R. A. (2014). Demand response in liberalized electricity markets: Analysis of aggregated load participation in the german balancing mechanism. *Energy*, 71:245–254.

Larsen, E. M. (2016). *Demand response in a market environment*. PhD thesis, Ph. D. dissertation.

Le Dréau, J. and Heiselberg, P. (2016). Energy flexibility of residential buildings using short term heat storage in the thermal mass. *Energy*, 111:991–1002.

Lee, S. (2019). COMPARING METHODS FOR CUSTOMER BASELINE LOAD ESTIMATION FOR RESIDENTIAL DEMAND RESPONSE IN SOUTH KOREA AND FRANCE: PREDICTIVE POWER AND POLICY IMPLICATIONS. *Chaire European Electricity Markets*, Working paper #39:51.

Lu, X., Li, K., Xu, H., Wang, F., Zhou, Z., and Zhang, Y. (2020). Fundamentals and business model for resource aggregator of demand response in electricity markets. *Energy*, 204:117885.

Mancini, F., Romano, S., Lo Basso, G., Cimaglia, J., and de Santoli, L. (2020). How the Italian Residential Sector Could Contribute to Load Flexibility in Demand Response Activities: A Methodology for Residential Clustering and Developing a Flexibility Strategy. *Energies*, 13(13):3359.

Mohajeryami, S., Doostan, M., Asadinejad, A., and Schwarz, P. (2016). Error analysis of customer baseline load (cbl) calculation methods for residential customers. *IEEE Transactions on Industry Applications*, 53(1):5–14.

Molderink, A., Bakker, V., Bosman, M. G., Hurink, J. L., and Smit, G. J. (2010). A three-step methodology to improve domestic energy efficiency. In *2010 Innovative Smart Grid Technologies (ISGT)*, pages 1–8. IEEE.

Müller, F. and Jansen, B. (2019). Large-scale demonstration of precise demand response provided by residential heat pumps. *Applied Energy*, 239:836–845.

Pina, A., Silva, C., and Ferrão, P. (2012). The impact of demand side management strategies in the penetration of renewable electricity. *Energy*, 41(1):128–137.

Prettico, G., Flammini, M., Andreadou, N., Vitiello, S., Fulli, G., and Masera, M. (2019). Distribution system operators observatory 2018. *Publications Office of the European Union*.

Srivastava, A., Van Passel, S., Kessels, R., Valkering, P., and Laes, E. (2020). Reducing winter peaks in electricity consumption: A choice experiment to structure demand response programs. *Energy Policy*, 137:111183.

Stamminger, R. and Anstett, V. (2013). The effect of variable electricity tariffs in the household on usage of household appliances. *Smart Grid and Renewable Energy, Vol. 4, No. 4, July 2013*.

Stede, J., Arnold, K., Dufter, C., Holtz, G., von Roon, S., and Richstein, J. C. (2020). The role of aggregators in facilitating industrial demand response: Evidence from germany. *Energy Policy*, 147:111893.

Sun, M., Strbac, G., Djapic, P., and Pudjianto, D. (2019a). Preheating Quantification for Smart Hybrid Heat Pumps Considering Uncertainty. *IEEE Transactions on Industrial Informatics*, 15(8):4753–4763.

Sun, M., Wang, Y., Teng, F., Ye, Y., Strbac, G., and Kang, C. (2019b). Clustering-Based Residential Baseline Estimation: A Probabilistic Perspective. *IEEE Transactions on Smart Grid*, 10(6):6014–6028.

United Nations Climate Change (2015). The paris agreement. `https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVII-7-d&chapter=27&clang=_en`.

Vallés, M., Bello, A., Reneses, J., and Frías, P. (2018). Probabilistic characterization of electricity consumer responsiveness to economic incentives. *Applied Energy*, 216:296–310.

Venizelou, V., Makrides, G., Efthymiou, V., and Georghiou, G. E. (2020). Methodology for deploying cost-optimum price-based demand side management for residential prosumers. *Renewable Energy*, 153:228–240.

Wang, F., Li, K., Liu, C., Mi, Z., Shafie-Khah, M., and Catalão, J. P. (2018). Synchronous pattern matching principle-based residential demand response baseline estimation: Mechanism analysis and approach description. *IEEE Transactions on Smart Grid*, 9(6):6972–6985.

Wang, X. and Tang, W. (2020). Analysis and evaluation of baseline manipulation in demand response programs. *arXiv preprint, arXiv:2011.10681*.

Wang, Y., Zhang, N., Tan, Y., Hong, T., Kirschen, D. S., and Kang, C. (2019). Combining Probabilistic Load Forecasts. *IEEE Transactions on Smart Grid*, 10(4):3664–3674.

Weng, Y., Yu, J., and Rajagopal, R. (2018). Probabilistic baseline estimation based on load patterns for better residential customer rewards. *International Journal of Electrical Power & Energy Systems*, 100:508–516.

Wi, Y.-M., Kim, J.-H., Joo, S.-K., Park, J.-B., and Oh, J.-C. (2009). Customer baseline load (CBL) calculation using exponential smoothing model with weather adjustment. In *2009 Transmission & Distribution Conference & Exposition: Asia and Pacific*, pages 1–4, Seoul, South Korea. IEEE.

Wijaya, T. K., Vasirani, M., and Aberer, K. (2014). When bias matters: An economic assessment of demand response baselines for residential customers. *IEEE Transactions on Smart Grid*, 5(4):1755–1763.

Wijbenga, J. P., MacDougall, P., Kamphuis, R., Sanberg, T., van den Noort, A., and Klaassen, E. (2014). *Multi-goal optimization in PowerMatching city: A smart living lab*. IEEE.

Wooldridge, J. M. (2020). *Introductory econometrics : a modern approach*. Cengage, Boston, MA, 7. edition. edition.

Xenergy (2002). Protocol development for demand response calculation: draft findings and recommendations, california energy commission (2002).

Zhang, Y., Chen, W., Xu, R., and Black, J. (2016). A Cluster-Based Method for Calculating Baselines for Residential Loads. *IEEE Transactions on Smart Grid*, 7(5):2368–2377.

ČEPS (2022). Setkání s poskytovateli a zájemci o poskytování svr – sekce energetický obchod. *Presented in an online format meeting with balancing services providers by ČEPS, Prague (16.03.2022)*.
Available for download at: `https://www.ceps.cz/cs/poskytovatele-pps` under "220316 - Setkání s poskytovateli.pdf".

Appendix

DR event days – Dates

Table 15: A. DR event days dates – Decrease Request

#	Decrease DR event days	weekday	#	Decrease DR event days	weekday
1	2021-12-11	Sat	35	2022-01-19	Wed
2	2021-12-12	Sun	36	2022-01-20	Thu
3	2021-12-13	Mon	37	2022-01-22	Sat
4	2021-12-14	Tue	38	2022-01-23	Sun
5	2021-12-15	Wed	39	2022-01-24	Mon
6	2021-12-16	Thu	40	2022-01-25	Tue
7	2021-12-18	Sat	41	2022-01-26	Wed
8	2021-12-19	Sun	42	2022-01-27	Thu
9	2021-12-20	Mon	43	2022-01-29	Sat
10	2021-12-21	Tue	44	2022-01-30	Sun
11	2021-12-22	Wed	45	2022-01-31	Mon
12	2021-12-23	Thu	46	2022-02-01	Tue
13	2021-12-25	Sat	47	2022-03-01	Tue
14	2021-12-26	Sun	48	2022-03-02	Wed
15	2021-12-27	Mon	49	2022-03-03	Thu
16	2021-12-28	Tue	50	2022-03-05	Sat
17	2021-12-29	Wed	51	2022-03-06	Sun
18	2021-12-30	Thu	52	2022-03-07	Mon
19	2022-01-01	Sat	53	2022-03-08	Tue
20	2022-01-02	Sun	54	2022-03-09	Wed
21	2022-01-03	Mon	55	2022-03-10	Thu
22	2022-01-04	Tue	56	2022-03-12	Sat
23	2022-01-05	Wed	57	2022-03-13	Sun
24	2022-01-06	Thu	58	2022-03-14	Mon
25	2022-01-08	Sat	59	2022-03-15	Tue
26	2022-01-09	Sun	60	2022-03-16	Wed
27	2022-01-10	Mon	61	2022-03-17	Thu
28	2022-01-11	Tue	62	2022-03-19	Sat
29	2022-01-12	Wed	63	2022-03-20	Sun
30	2022-01-13	Thu	64	2022-03-21	Mon
31	2022-01-15	Sat	65	2022-03-22	Tue
32	2022-01-16	Sun	66	2022-03-23	Wed
33	2022-01-17	Mon	67	2022-03-24	Thu
34	2022-01-18	Tue	68	2022-03-26	Sat

Table 16: DR event days dates – Increase Request

#	Increase DR event days	weekday	#	Increase DR event days	weekday
1	2021-12-11	Sat	24	2022-01-19	Wed
2	2021-12-13	Mon	25	2022-01-22	Sat
3	2021-12-14	Tue	26	2022-01-24	Mon
4	2021-12-15	Wed	27	2022-01-25	Tue
5	2021-12-18	Sat	28	2022-01-26	Wed
6	2021-12-20	Mon	29	2022-01-29	Sat
7	2021-12-21	Tue	30	2022-01-31	Mon
8	2021-12-22	Wed	31	2022-02-01	Tue
9	2021-12-25	Sat	32	2022-03-01	Tue
10	2021-12-27	Mon	33	2022-03-02	Wed
11	2021-12-28	Tue	34	2022-03-05	Sat
12	2021-12-29	Wed	35	2022-03-07	Mon
13	2022-01-01	Sat	36	2022-03-08	Tue
14	2022-01-03	Mon	37	2022-03-09	Wed
15	2022-01-04	Tue	38	2022-03-12	Sat
16	2022-01-05	Wed	39	2022-03-14	Mon
17	2022-01-08	Sat	40	2022-03-15	Tue
18	2022-01-10	Mon	41	2022-03-16	Wed
19	2022-01-11	Tue	42	2022-03-19	Sat
20	2022-01-12	Wed	43	2022-03-21	Mon
21	2022-01-15	Sat	44	2022-03-22	Tue
22	2022-01-17	Mon	45	2022-03-23	Wed
23	2022-01-18	Tue	46	2022-03-26	Sat

B. Days acceptable for estimation - Dates

Table 17: Acceptable dates for **weekday** baseline estimation

#	dates	weekday
1	2021-12-17	Fri
2	2021-12-24	Fri
3	2021-12-31	Fri
4	2022-01-07	Fri
5	2022-01-14	Fri
6	2022-01-21	Fri
7	2022-01-28	Fri
8	2022-02-02	Wed
9	2022-02-03	Thu
10	2022-02-04	Fri
11	2022-02-07	Mon
12	2022-02-08	Tue
13	2022-02-09	Wed
14	2022-02-10	Thu
15	2022-02-11	Fri
16	2022-02-14	Mon
17	2022-02-15	Tue
18	2022-02-16	Wed
19	2022-02-17	Thu
20	2022-02-18	Fri
21	2022-02-21	Mon
22	2022-02-22	Tue
23	2022-02-23	Wed
24	2022-02-24	Thu
25	2022-02-25	Fri
26	2022-02-28	Mon
27	2022-03-04	Fri
28	2022-03-11	Fri
29	2022-03-18	Fri
30	2022-03-25	Fri

Table 18: Acceptable dates for **weekend** baseline estimation

	dates	weekday
1	2022-02-05	Sat
2	2022-02-06	Sun
3	2022-02-12	Sat
4	2022-02-13	Sun
5	2022-02-19	Sat
6	2022-02-20	Sun
7	2022-02-26	Sat
8	2022-02-27	Sun

C. Regression Method – Heteroscedasticity Testing

In section 5.3 we present a linear and quadratic model, using OLS to estimate our coefficients. As noted in the main body of the paper, we consider the presence of heteroscedasticity, which would mean we would have to use a GLS or FGLS estimator. However, after testing for heteroscedasticity (both numerically and graphically) we come to the conclusion that our data is borderline homoscedastic and that specifying an other-than-OLS model would not add much precision. The figure below shows the plotted residuals againts the fitted values, along with a Q-Q normality plot and other characteristics. The Breusch-Pagan test statistic (BP) was 2.8826 with a p-value of 0.08954, therefore, we fail to reject the null of homoscedasticity at a 8.9% significance level. We acknowledge this doesn't follow the standard rule of thumb (5 % significance level), however for reasons discussed in the paper, we don't see a need to deal with this any further.

D. Performance metrics for non-DR weekends example

Table 19: Performance Metrics for Weekend non-DR Estimation

Method	MAE (kWh)	Bias (kWh)
Mid 2of4	1.0500	-0.1832
High 2of3	1.0401	-0.1042
Low 2of3	1.0440	-0.4234
Exp. MA	1.0325	-0.1505
Exp. Smooth	0.9595	-0.1700
Reg. Linear	1.0016	-0.0348
Reg. Quadratic	1.0497	-0.0316

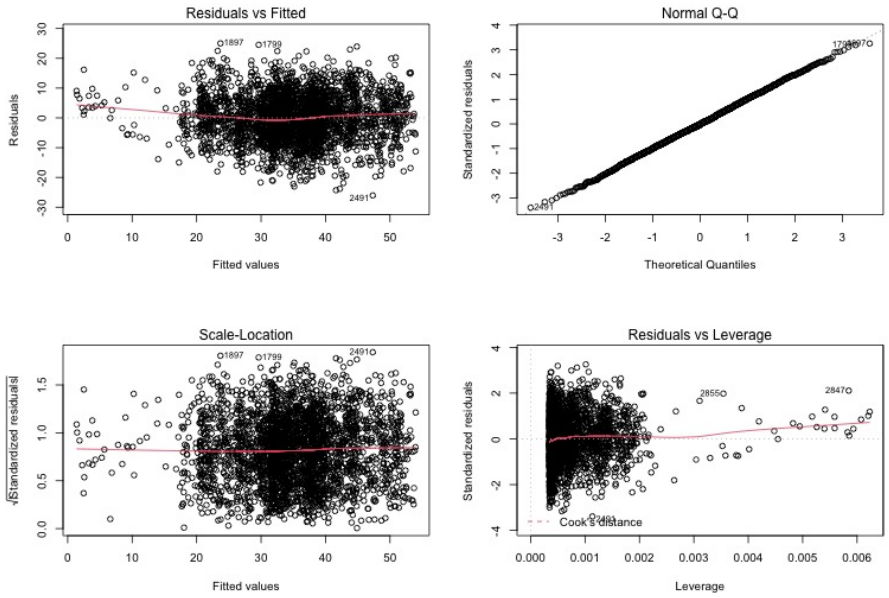


Figure 17: Heteroscedasticity and Normality Plots.

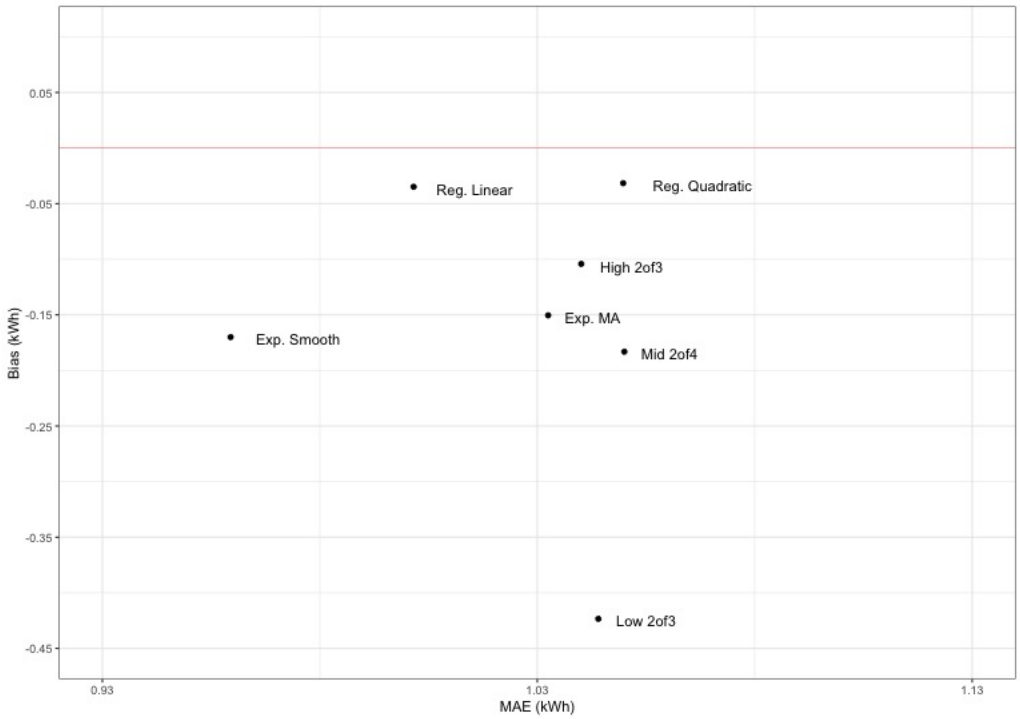


Figure 18: Performance metrics for Weekend non-DR Estimation

E. Quantification of DR Scheme

Table 20: Quantification for All Estimated Methods – Only DR Event Duration

Method & Request Type	Mon	Tue	Wed	Thu	Method Sum
Mid 4of6 (D)	9.45	27.42	24.54	18.65	80.05
Mid 4of6 (I)	-50.80	-70.27	-19.33	0.00	-140.40
sum (kWh)	-41.35	-42.85	5.21	18.65	-60.35
Mid 4of6 Adj. (D)	20.67	17.79	28.69	32.34	99.49
Mid 4of6 Adj. (I)	32.25	-26.47	-21.29	0.00	-15.50
sum (kWh)	52.92	-8.68	7.40	32.34	83.98
High 5of10 (D)	16.72	44.37	39.89	29.72	130.70
High 5of10 (I)	-36.75	-43.56	-4.73	0.00	-85.03
sum (kWh)	-20.03	0.81	35.16	29.72	45.66
High 5of10 Adj. (D)	8.38	10.01	10.62	22.67	51.69
High 5of10 Adj. (I)	-18.07	-40.53	-25.88	0.00	-84.49
sum (kWh)	-9.70	-30.52	-15.26	22.67	-32.80
Low 5of10 (D)	-1.96	-4.62	14.87	8.46	16.75
Low 5of10 (I)	-65.30	-104.75	-34.50	0.00	-204.55
sum (kWh)	-67.26	-109.37	-19.63	8.46	-187.80
Low 5of10 Adj. (D)	19.73	15.02	49.21	29.84	113.80
Low 5of10 Adj. (I)	114.28	116.97	3.66	0.00	234.91
sum (kWh)	134.01	132.00	52.87	29.84	348.72
Exp. Moving Average (D)	8.75	23.12	29.23	20.52	81.61
Exp. Moving Average (I)	-45.74	-62.49	-15.24	0.00	-123.47
sum (kWh)	-36.99	-39.38	13.98	20.52	-41.87
Exp. Moving Average Adj. (D)	8.90	3.43	9.39	20.52	42.25
Exp. Moving Average Adj. (I)	-24.65	-45.66	-27.32	0.00	-97.62
sum (kWh)	-15.74	-42.23	-17.93	20.52	-55.37
Exp. Smoothing (D)	7.64	20.75	28.10	19.21	75.70
Exp. Smoothing (I)	-47.32	-65.23	-16.93	0.00	-129.49
sum (kWh)	-39.68	-44.49	11.17	19.21	-53.79
Exp. Smoothing Adj. (D)	9.06	3.47	10.10	21.16	43.79
Exp. Smoothing Adj. (I)	-26.00	-46.54	-27.57	0.00	-100.10
sum (kWh)	-16.94	-43.07	-17.47	21.16	-56.32
Linear Regression (D)	20.09	47.63	19.90	18.57	106.19
Linear Regression (I)	-22.38	-39.40	-30.54	0.00	-92.33
sum (kWh)	-2.29	8.23	-10.65	18.57	13.86
Linear Regression Adj. (D)	10.30	13.39	21.55	26.57	71.81
Linear Regression Adj. (I)	-25.34	-44.39	-39.24	0.00	-108.97
sum (kWh)	-15.04	-31.00	-17.69	26.57	-37.17
Quadratic Regression (D)	19.31	46.68	20.55	18.70	105.25
Quadratic Regression (I)	-20.45	-38.76	-30.04	0.00	-89.24
sum (kWh)	-1.13	7.92	-9.49	18.70	16.01
Quadratic Regression Adj. (D)	10.07	13.99	9.13	23.27	56.46
Quadratic Regression Adj. (I)	-25.18	-42.82	-40.20	0.00	-108.20
sum (kWh)	-15.11	-28.83	-31.07	23.27	-51.73

F. Synchronous Pattern Matching - Proposed Method

Synchronous Pattern Matching is a method proposed by Wang et al. in 2018, where method matches each DR participant to the most similar group of participants in a control cluster. When this has been done, an estimate of the DR participants baseline is made based on the concurrent load data of the control cluster. Unfortunately, since this method requires at least a similar amount of DR participants and controls (preferably more controls than DR participants for clusters to be created) and our data only having 5 controls compared to 22 DR-participants, it would be unfeasible to use this method with our dataset. Nevertheless, we propose this method for possible further research. This following section is based on Wang et al. (2018).

The method divides the dataset into M CONTROL groups and N DR groups. $D = \{d|d = 1, 2, ..., D\}$ is defined as the set of DR event days and $T = \{t|t = 1, 2, ..., T\}$ as the set of timeslots for a DR event day. The CONTROL group customers are separated based on their load profiles (LP) on the DR event day based on an iterative process. A cluster's centroid is derived by computing the average of every data point found in the cluster. The aim of K-means is to minimize the sum of squared error between the CONTROL customers load curves and the cluster centroids over all clusters K for each DR event day *d*. This is defined in the formula below.

$$minf = \sum_{k=1}^{K} \sum_{m=1, m \in C_k} L_{m,d} - C_{k,d} \forall k = 1, ..., K$$

$L_{m,d} = [l^1_{m,d}, l^2_{m,d}, ..., l^T_{m,d}]$ is the actual load curve of control customer m.
$C_{k,d} = [c^1_{k,d}, c^2_{k,d}, ..., c^T_{k,d}]$ is the cluster centroid.

To assess clustering performance, the Davies-Bouldin index (DBI) and Ratio of within Cluster Sum of Squares to Between Cluster Variation (WCBCR) are used. A key component of this method is the *similarity* metric. For each DR event day, each DR participant should be matched to one of the obtained K clusters in the CONTROL group. SPM here refers to using the very DR event day, therefore abstracting from the need of any historical data out of the DR event day. Wang et al. represent the DR event time period as δ where $\delta = \{\delta_s, \delta_s + 1, ..., \delta_e\}, \delta_s$ being the start and $\delta_e < T$ being the end. For each DR day $d \in D$, the actual load data of DR participant n before and after DR event can be utilized to perform the SPM based CBL estimation, which is defined as load curve

segments (LCSs), denoted by

$$LCS_{n,d}^{before} = [l_{n,d}^{1}, l_{n,d}^{2}, ..., l_{n,d}^{\delta_s - 1}]$$

and

$$LCS_{n,d}^{after} = [l_{n,d}^{\delta_e + 1}, l_{n,d}^{\delta_e + 2}, ..., l_{n,d}^{\delta_e + T}]$$

Wang et al. define the cluster centroid k load data before and after the DR evenet similarly. These are called cluster centroid segments (CCSs).

$$CCS_{k,d}^{before} = [c_{k,d}^{1}, c_{k,d}^{2}, ..., c_{k,d}^{\delta_s - 1}]$$

and

$$CCS_{k,d}^{after} = [c_{k,d}^{\delta_e + 1}, c_{k,d}^{\delta_e + 2}, ..., c_{k,d}^{\delta_e + T}]$$

Then, the similarity between vectors x and y is given by the formula below

$$S(x, y) = \frac{1}{dis(x, y)}$$

$dis(x, y)$ is the distance between the two vectors. A common distance metric such as Euclidean distance can be used. The larger the value $S(x, y)$ obtains the more similar the two vectors are. The SPM is done based on the similarity between the LCSs and CCSs. You calculate the similarity between $LCS_{n,d}^{before}$ and each $CCS_{k,d}^{before}$ (for $k = 1, 2, ..., K$) denoted as $S(LCS_{n,d}^{before}, CCS_{k,d}^{before})$, for each DR customer n . Additionally, the similarity between $LCS_{n,d}^{after}$ and each $CCS_{k,d}^{after}$ (for $k = 1, 2, ..., K$), denoted as $S(LCS_{n,d}^{after}, CCS_{k,d}^{after})$ is calculated. The similarity between each DR customer and cluster C_k (for $k = 1, 2, ..., K$)

can then be expressed as the formula below.

$$S(DRcustomer_n, C_k) = S(LCS_{n,d}^{before}, CCS_{k,d}^{before}) + S(LCS_{n,d}^{after}, CCS_{k,d}^{after})$$

By this equation the DR participant will be matched to the cluster that shows the maximum similarity with the DR participant.

Finally, Wang et al. estimate the CBL for each DR participant. For this the authors use optimized weight combination. First, they find all the control customers, belonging to cluster k, to which a certain DR participant n has been assigned to. These are indexed as follows, $I_k = \{1, 2, ..., M_k\}$ where M_k is the number of CONTROL customers in cluster k. Since the customers in the CONTROL cluster have load profiles similar to the DR participant n, the load of each CONTROL customer n can be seen as individual baseline estimations for DR participant n. Looking for inspiration in Bates and Granger (1969), Wang et al. (2018) decide to combine the results of multiple different forecast models to effectively improve the forecasting accuracy. Therefore, a "combination estimation model" is created, which combines the baselines of all customers in cluster k in order to estimate the baseline for DR participant n. The formula of the above looks as follows.

$$b_{n,d}^t = f_{i \in I_k}(l_i(d,t), \forall n \in C_k, t = \delta_s, ..., \delta_e$$

$b_{n,d}^t$ is the estimated baseline for said DR participant and $f(\cdot)$ is a function that maps the load of the CONTROL customers to the baseline of the DR participant. Wang et al. (2018) Wang et al. (2018) use the following linear combination for the mapping function,

$$b_{n,d}^t = \sum_{i=1}^{M_k} w_i l_{i,d}^t, t = \delta_s, ..., \delta_e$$

w_i is the weight of the i^{th} individual estimation model, which is corresponding to the i^{th} CONTROL customer in cluster k. At this point the method hits one of the key issues. That is, finding an optimal set of non-negative weights $W = [w_1, w_2, ..., w_{M_k}]$ such that the estimated CBL as close to the actual load as possible.
Because the actual baseline load is unknown in reality, only the load data outside of the

DR event duration can be used when determining these weights. Denoting this "outside of DR event" time period by $\varepsilon = \{1, 2, ..., \delta_{s-1} U \{\delta_{e+1}, \delta_{e+2}, ..., T\}$ and $T = \delta U \varepsilon$. $e_{it} = l^t_{n,d} - l^t_{i,d}, t \in \varepsilon$ denotes the error of the ith individual estimation model at timeslot t.

An estimation error vector ε can formed by all of the individual estimation models, which is denoted by $e_i = [e_{i1}, e_{i2}, ..., e_{i|\varepsilon|}]^T$, where denotes $|\varepsilon|$ the amount of timeslots in time period ε . The error of the combined estimation model at the time slot t can be calculated as follows.

$$e_t = b^{t*}_{n,d} - b^t_{n,d} = l^t_{n,d} - \sum_{i=1}^{M_k} w_i l^t_{i,d} = \sum_{i=1}^{m_k} w_i e_{it}$$

$b^{t*}_{n,d}$ is the actual baseline, which is the same as load $l^t_{n,d}$ during the same time period without the DR event.

Next, the linear combination is formulated as an optimization model where the goal is to find the optimal weights to minimize the sum of squared errors. It is formulated followingly.

$$minJ = \sum_{t=1}^{|\varepsilon|} e^2_{it} = \sum_{t=1}^{|\varepsilon|} \sum_{i=1}^{M_k} \sum_{j=1}^{M_k} w_i e_{it} w_j e_{jt}$$

$$s.t. \begin{cases} \sum_{i=1}^{M_k} w_i = 1 \\ w_i \geq 0, i = 1, 2, ..., M_k \end{cases}$$

To offer this model in matrix form, Wang et al. define a square matrix named error information matrix with the size of $M_k x M_k$ denoted as $E_{(M_k)} = (E_{ij})_{M_k x M_k}$. The element in this matrix is defined below.

$$\begin{cases} E_{ij} = e^T e_i = \sum_{t=1}^{|\delta|} e^2_{it} \\ E_{ij} = e^T e_j = \sum_{t=1}^{|\delta|} e_{it} e_{jt} \end{cases}$$

Then to rewrite the optimization in above a vector which has all it's element values equal to 1 has to be defined. It is denoted as $R = [1, 1, ..., 1]^T$. Then, the optimisation problem

from above can be written as follows.

$$minJ = W^T EW$$
$$s.t. \begin{cases} R^T W = 1 \\ W \geq 0 \end{cases}$$

To solve this the Langrange multiplier method used, and the optimal weights can be obtained from below.

$$W = \frac{E^{-1}R}{R^T E^{-1} R}$$

When optimal weights have been obtained the baseline is possible to be obtained properly for the DR participant i based on control cluster k.