newTRENDs

Impact assessment of household-level behavioral interventions via smart-meter data

Deliverable 4.2

Grant agreement	No. 893311	Acronym	newTRENDs
Full title	New Trends in Energy Demand Modeling		
Topic	LC-SC3-EE-14-2018-2019-2020		
Funding scheme	Horizon 2020, RIA – Research and Innovation Action		
Start date	September 2020	Duration	36 Months
Project website	https://newtrends2020.eu/		
Project coordinator	Fraunhofer ISI		
Deliverable	4.2 - Impact assessment of household-level behavioral interventions via smart-meter data		
Work package	4		
Date of Delivery	Contractual 11/21	Actual	04/22
Status	Draft		
Nature	Report	Dissemination level	Public
Lead beneficiary	Politecnico di Milano (POLIMI)		
Responsible author	Giacomo Marangoni (POLIMI)		
Contributors	Irene Malvestio (POLIMI) Maksymilian Kochanski (RIC) Katarzyna Korczak (RIC)		
Reviewer(s)	Songmin Yu (Fraunhofer ISI) Martin Jakob, Joachim Bagemihl (TEP)		
Keywords	Smart meters, impact assessment, machine learning for causal inference, behavioral intervention, energy savings, energy demand, behavioral change		

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under grant agreement no. 893311.

Executive Summary

Energy-demand options for reducing household power consumption have been discussed as valuable and cost-effective options complementary to the more expensive energy supply measures required to reduce the carbon footprint of the residential energy sector. While power utilities are exploring opportunities for energy conservation coming from consumers' behavioral change, new societal trends, like digitalization, and socio-economic contexts, as the one following the COVID-19 pandemic, have profoundly affected the way households consume energy, and not necessarily steering consumption towards lower values.

This report overviews the opportunities offered by both traditional and novel machine learning techniques to assess the causal outcome of large-scale behavioral interventions affecting power consumption. By applying these methods to the large smart data ensembles collected in natural experimental contexts, like those set up by utilities or even those naturally arisen by the progressive application of lockdown policies during the COVID-19 pandemic, it is possible to explore a variety of response models to different behavioral levers, while trying to unpack the heterogeneity and to guide the policy discussion.

Different feedback mechanisms have been devised with varying level of success in affecting consumption. The causal effect of such interventions has been traditionally estimated with econometric techniques. More recently, thanks to the availability of large data ensemble from smart meters, machine learning brings new opportunities for causal inference analyses.

This work reviews available machine learning techniques that can be used to leverage on smart meter datasets for impact evaluation. Concepts like forecasting, clustering, explainable machine learning, and causal forests are presented for this purpose and their benefits and limits emphasized.

Three case studies corresponding to natural and artificial behavioral interventions monitored via smart meters in Italy and Poland showcase a variety of traditional and novel techniques used to analyze the corresponding smart meter datasets and uncover patterns of household-behavioral changes in power consumption associable with those interventions.

Overall, novel large-scale data-driven assessments of behavioral intervention suggest energy savings of few percentage points. Nonetheless high heterogeneity emerges from the data. Machine learning can help to better understand this heterogeneity. In general, more data and experiments are needed to further refine the match between different classes of households and the most effective behavioral intervention, as well as to scale the insights to other regional contexts.

Table of content

List of figures

Table of tables

1. Introduction

Europe has set ambitious targets for energy transition and decarbonization, with the aim to develop a sustainable, low-carbon society. Existing policy packages, such as the 2030 energy and climate strategy and the EU Green Deal, have emphasized a variety of proposed tools. Energy-demand options have been discussed as valuable and cost-effective options complementary to the more expensive energy supply measures required to reduce the carbon footprint of the residential energy sector. Power utilities are thus exploring opportunities for energy conservation coming from consumers' behavioral change. At the same time, new trends, like digitalization, and socio-economic contexts, as the one following the COVID-19 pandemic, have profoundly affected the way households consume energy, and not necessarily steering consumption towards lower values.

Avoiding inefficient behaviors and appliances in households (HHs) have multiple benefits. On one hand, households can reduce their energy bills by reducing or shifting consumption. On the other, power providers can avoid potentially costly and carbon-intensive on-peak production, as well as investments in networks and plants to sustain otherwise higher future peak demands. Thanks to the increasing deployment of smart meters, i.e. electronic devices that record consumption of electric energy with high frequency for monitoring and billing purposes, new possibilities for consumers' engagement and data-driven policies arise. Smart meters are indeed being deployed at a very fast rate, with almost 225 million devices expected to be operational in the EU in 2024. Such technology has the capability of recording and communicating energy consumption levels at a high-temporal resolution (e.g., 15 minutes or less), thus enabling consumers with an on-hand grasp of their energy bills.

Policy makers have promoted different types of behavioral programs, claiming their high potential and cost effectiveness for energy conservation. Combined with the digital support of smart meters, it is argued that users' behavior can be nudged towards better environmental choices by overcoming limiting decision factors such as limited attention, present bias and limited salience (Allcott and Mullainathan 2010). However, critics remain due to mixed empirical findings (Andor et al. 2022).

This report overviews the opportunities offered by both traditional and novel machine learning techniques to assess the causal outcome of large-scale behavioral interventions. By applying these methods to the large smart data ensembles collected in natural experimental contexts, like those set up by utilities or even those naturally arisen by the progressive application of lockdown policies during the COVID-19 pandemic, it is possible to explore a variety of response models to different behavioral levers, while trying to unpack the emerging heterogeneity.

1.1 Policy (evaluation) context

Under the Directive (EU) 2018/2002 ("Directive (EU) 2018/2002 of the European Parliament and of the Council of 11 December 2018 Amending Directive 2012/27/EU on Energy Efficiency, OJ l 328" 21.12.2018), the European Union set an energy efficiency target for reducing energy consumption in 2030 by at least 32.5% compared to 2007 consumption projections for 2030. To this end, a wide range of technological, economic, and behavioral changes need to be implemented across the Union. The regulatory requirements concerning the latter type of changes are provided by art. 12 of the Energy Efficiency Directive (("Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on Energy Efficiency, Amending Directives 2009/125/EC and 2010/30/EU and Repealing Directives 2004/8/EC and 2006/32/EC, OJ l 315" 14.11.2012) 2012/27/EU, further: EED). It obliges the Member States to implement instruments and policies to promote behavioral change of consumers towards an efficient use of energy. The Directive enumerates the following policy instruments in this area:

- fiscal incentives;
- access to finance, grants, or subsidies;
- information provision;
- exemplary projects;
- workplace activities.

The cost-effectiveness of these instruments may vary, but it is important that these are carefully evaluated before making rational policy choices. In this context, the valuation of environmental effects (here: energy savings) is indispensable for policy making and should be based on well-established methods (Mickwitz 2003). Annex V to the EED specifies the following possible methods for calculating energy savings[1]:

- deemed savings – referring to results of previous independently monitored energy improvements in similar installations;
- metered savings – referring to the measured reduction in energy use;
- scaled savings – referring to engineering estimates (independent benchmarks);

1 The listed methods concern behavior measures other than those arising from taxation. With regard to calculating energy savings arising from taxation measures, Annex V to the EED requires calculation methods to ensure that: (1) Only energy savings from taxation measures exceeding the minimum levels of taxation applicable to fuels as required in Council Directive 2003/96/EC or 2006/112/EC are considered; (2) price elasticities shall represent the responsiveness of energy demand to price changes; (3) the energy savings from accompanying taxation policy instruments shall be accounted separately. Calculating energy savings arising from taxation measures is not within the scope of this Deliverable.

- surveyed savings – referring to changes in consumer behavior.

Calculating energy savings from large-scale behavior change entails several challenges. First, the materiality of a change in behavior is more problematic to show than in the case-technical changes, such as a replacement of appliances for more energy-efficient ones. Second, behavior changes can be easily reversed, so determining the lifetime of the measures and the amount of savings over time is inherently linked with a high degree of uncertainty. In this perspective, the European Commission (EC) stresses that behavioral measures require specific evaluation methods (Appendix VI to the Commission Recommendation (EU) 2019/1658 ("Commission Recommendation (EU) 2019/1658 of 25 September 2019 on Transposing the Energy Savings Obligations Under the Energy Efficiency Directive, OJ l 275" 28.10.2019)). The EC recommends Member States to use one of the three approaches:

- randomized controlled trials;
- quasi-experimental approaches;
- metering or monitoring energy consumption.

They all have the same objective, which is the calculation of energy savings, but each approach offers different advantages and is linked with different limitations (Table 1). Apart from different costs and level of operational difficulty in implementation, these methods entail also diverse levels of risk of confounding (mixing or blurring of the effects), i.e. suggesting an association between the behavioral measure and energy savings where none exists or masks a true association (JK, C, and D 2018).

Table 1: Comparison of methods for evaluating energy savings from behavioral measures according to the Commission Recommendation (EU) 2019/1658.

Method / Method characteristic	Randomized controlled trials	Quasi-experimental approach	Metering or monitoring energy consumption
Random allocation of participants into groups: • a treatment group that experiences an intervention • a control group that does not receive the intervention	Yes	No. Treatment group is compared with a comparison group, which is not chosen randomly from the same population as treatment group. The allocation of participants to a treatment group and comparison group could be based on, for instance, self-selection, or according to program owner decisions.	No. Energy savings are evaluated by metering or monitoring the participants' energy consumption before and after the intervention.
Level of evidence on the impacts of the behavioral measures	Highest – most definitive casual inference	Medium	Lowest
Level of risk of confounding	Not likely [2]	Likely	Very likely
Level of difficulty in operational implementation	Highest	Medium-high	Lowest

In all methods, smart metering data is highly useful. It offers a less expensive, faster, and more reliable calculation of energy savings from behavioral measures, as compared to methods based on manually collected energy consumption data of lower temporal granularity. Smart meters offer also a

2 In some cases confounding can occur even in a randomized control trial, e.g. due to knowledge on participant allocation to different study groups by the researchers (Manson et al. 2016).

higher precision of energy savings calculation, which is of paramount importance in view of the relatively low reductions of energy use due to such measures – typically less than 3% (Stewart, Todd, and Kurnik 2020). Data from smart meters can also contribute to a better understanding of the common barriers for promotion of energy efficient behaviors in households. In particular, smart meter data can provide reliable evidence on:

- effectiveness of specific components of behavioral measures – through assessing impacts of differed treatments on household energy consumption;
- lifetime of energy savings – through assessing the length of time during which a behavioral measure generates energy savings;
- persistence of energy savings – through assessing the change in savings throughout their lifetime;
- effectiveness of behavioral measures on energy consumption of different households – though assessing impacts of the treatment on households with different socio-economic or technical characteristics.

1.2 Overview

Section 2 offers a literature review of behavioral interventions studies where households have been incentivized to change their power consumption trends to save energy. Different feedback mechanisms have been devised with varying level of success in affecting consumption. The causal effect of such interventions has been traditionally estimated with econometric techniques. More recently, thanks to the availability of large data ensemble from smart meters, machine learning brings new opportunities for causal inference analyses.

Section 3 goes into greater detail regarding available machine learning techniques that can be used to explore smart meter datasets for impact evaluation. Within this perspective, concepts like forecasting, clustering, explainable machine learning, and causal forests are presented for this purpose and their benefits emphasized.

Section 4, 5 and 6 present three case studies where a variety of traditional and novel techniques are showcased in uncovering patterns of household-behavioral changes in power consumption associable with both natural and artificial interventions.

Finally, Section 7 concludes the report with a discussion on the relevance of this work and illustration of future directions.

2. Literature review

2.1 Exogenous intervention due to COVID-19

The restrictions in mobility that has followed the COVID-19 pandemic have influenced not only the level of the electricity demand in Europe, but also consumption patterns (Narajewski and Ziel 2020), which reflect the various rules in places on different European countries (Bahmanyar, Estebsari, and Ernst 2020). A study of smart meters data from 280 homes in Cornwall (Menneer et al. 2021), UK, found increase in gas, water and electricity usage, with electricity consumption that shift later in the day, reflecting people getting up later and spending more time at home. Other studies in the literature (Menneer et al. 2021; Cuerdo-Vilches, Navas-Martín, and Oteiza 2021) underline how the economic impact of the increased consumption affects people in socioeconomically disadvantaged areas, adding to the other inequitable impacts of the virus. During the first lockdown in 2020, economic activities have severely reduced, and this reduction is mirrored with a reduction in electricity market consumption (Fezzi and Fanghella 2020; Soava et al. 2021).

COVID lockdowns provide a set of natural experiments to explore the effects on electricity consumption of work and study-from-home. In Li et al. (2021) data from 390 apartments in New York City (NYC) were analyzed with regression models to find how COVID-19 cases together with the weather influence power consumption loads. With Monte Carlo simulations it was possible to estimate how the peak demand would dramatically increase (almost double up) if future emergencies would lead to a stay-at-home order in a similar region to NYC, during the warmest months.

Work and study-from-home are public responses adopted to face the COVID-19 pandemic emergency, and their effects on consumption are worth being analyzed in themselves, considering that other lockdown may be taken place in the future in response to COVID-19, other pandemics, or other crisis. At the same time teleworking is not only a response to a pandemic, but it is also perceived as a more sustainable mode of working (for those jobs that are compatible) as it may be beneficial for employee lifestyle, it reduces costs for the companies, and it should have environmental benefits. For example, in Tenailleau et al. (2021), the authors find that in the medium-sized European city of Besançon (France), a 1% increase in teleworking leads to an average reduction in emission of −0.42%. The assessment of teleworking influence on energy consumption needs to take into account the increased usage in the household, as well as consumption reduction in the office, reduction of commuting, and all the complex rebound effects that a shift in the location of workers may imply (O'Brien and Aliabadi 2020). The literature shows no consensus on the estimates of energy saving due to teleworking: some studies point to a net saving, while others claim a more modest effect when non-work travel or home energy use are also considered (Hook et al. 2020).

Among other methods, clustering analysis has been applied to find the effect of COVID-19 on consumption. In García et al. (2021) the authors consider consumption of both residential and non-residential consumers from the town of Manzanilla (Huelva, Spain), to identify which categories increased/decreased their consumption during the lockdown. While residential consumption increased by about 15% during full lockdown and 7.5% during the reopening period, for non-residential consumers, despite an overall decrease in consumption, five different consumption profiles were found. Clustering techniques are applied also in Abdeen et al. (2021) for the study of different consumption behaviors before and during COVID lockdown in Ottawa, Canada. Authors estimated that electricity use of homes for cooling is not significantly affected by COVID-induced behaviors. Considering average electricity daily profiles, authors find that there are differences among different months, seasons, and day types in the lockdown influence over electricity consumption. In section 4.1of this report, we apply clustering methods to unravel the change in power consumption patterns during the first COVID-19 lockdown in Italy.

2.2 Behavioral interventions

People's perceptions of energy consumption can have significant discrepancies with reality (Attari et al. 2010). In particular, low-energy activities tend to be overestimated and high-energy activities tend to be largely underestimated. This justifies the exploration of behavioral interventions, or nudges, to promote lower energy use, more environmentally friendly decisions, and more generally virtuous behaviors (Sunstein 2021). So far studies have mostly looked at empirical evidence for social comparison interventions (Allcott 2011) and information feedback type of programs (Andor and Fels 2018).

2.2.1 Social feedback

In the last decade, electricity providers have been sending Home Energy Reports (HERs) providing a social comparison feedback to customers to encourage energy savings. HERs were first popularized by Opower in the USA beginning in 2008, based on the results of a field experiment that showed how reports are effective in promoting energy conservation (Schultz et al. 2007). HERs typically consist of a descriptive part, in which household consumption is compared with those of neighbors, or families with similar characteristics, while the second part is an injunctive feedback with social approval for energy saving (Bonan et al. 2020).

A large field of literature is dedicated to empirically estimate the effects of nudges in encouraging energy conservation. The first seminal works on HERs focus on the effectiveness of the Opower program in the USA, which consists on repeatedly sending HERs to more than six million households throughout the country. Allcott and Rogers (Allcott and Rogers 2014) are among the first to study the outcome of the program. Using linear regression, they find a reduction in electricity consumption in the days after the reception (until around 10 days), with a subsequent decrease in savings. After more than four reports, the effect of the individual report is weaker, but overall, the repetition of the treatment leads to a consistent reduction in consumption.

Multiple seminal studies (Allcott 2011) on the effects of Opower HERs in the USA show a reduction in electricity consumption of around 2% or more: 211 independent RCTs conducted in the USA by Opower find a reduction between 0.81% and 2.55% across states (Jachimowicz et al. 2018), and the electricity savings are confirmed not only by the Opower program (Henry, Ferraro, and Kontoleon 2019).

However, these results are context-dependent: they may be subject to site selection bias (i.e. the probability of adopting or evaluating a program is not independent from its impacts (Allcott 2015)), and in general for European countries the effect size of similar HER programs are much lower (i.e. 0.7% of reduction was found in Germany (Mark A. Andor et al. 2020)).

2.2.2 Consumption feedback

Programs providing more frequent or salient consumption information feedback have been also a subject of field experiments. On one hand, information feedback seems to strengthen the price elasticity of demand, making interventions like dynamic pricing more effective (Jessoe and Rapson 2014). On the other, it supposedly induces electricity savings (Schleich et al. 2013; Attari et al. 2014; Lynham et al. 2016). This result is driven both by the process of learning about the energy consumption of different activities as well as of reminding of one's energy use, with possibly the former being a stronger driver (Lynham et al. 2016).

Regarding relative average consumption reductions, two papers performed a meta-analysis of the many available studies (Karlin, Zinger, and Ford 2015; McKerracher and Torriti 2013). We refer the readers to the references therein for an exhaustive coverage of the literature on the topic. Karlin, Zinger, and Ford (2015) consider 42 studies with an effect size on power consumption ranging from -8% to above 20%. Their meta-model estimates an overall reduction between 4% and 12%, depending on different aggregation schemes. Several treatment variables moderate this relationship, including frequency of feedback, medium, comparison message, duration, and combination with other interventions (e.g., goal, incentive). Feedback turns out to be more effective especially when combined with goal-setting or external incentive interventions, and is somewhat brief (e.g., less than 3 months) or quite long (e.g., longer than 1 year). McKerracher and Torriti (2013) come up with more conservative estimates in their meta-analysis. Especially focusing on the lower estimates of more recent studies with larger sample sizes and more representative sample selection and recruitment methods, the authors argue that a realistic, large-scale conservation effect from feedback is in the range of 3-5%.

These estimates are significantly lower than what suggested in earlier studies, but are supposed to be more robust, given the very large sample sizes of recent trials. It should be remarked that even if relative reductions do not appear to be large, they are in fact considered very significant from the industry view point.

If these results can be generalized, given the large share of energy consumed in the residential and commercial sector, they can go a long way in improving energy efficiency, especially in the light of the very low-price elasticities observed in residential energy. Among available types of feedback, real-time household-level feedback seems to be the most effective in overcoming imperfect information and inattention biases, with potentially large conservation effects (Tiefenbeck et al. 2016).

As technology becomes cheaper and consumption data more readily available, interest in high-frequency feedback has been rising. Houde et al. (2013) takes advantage of hourly data from a long-lasting real-time feedback field experiment and look also into time-of-day reduction effects and persistence of the effect over time. In this case, access to feedback leads to an average reduction in household electricity consumption of 5.7%, persisting for up to four weeks. Another experiment combined in-home displays for real-time feedback with interventions involving moral suasion and economic incentives (Ito, Ida, and Tanaka 2018). Moral suasion led to 8% short-run power usage reductions, while economic incentives doubled reductions and made them more persistent, supporting the synergy between information and dynamic pricing (Ito, Ida, and Tanaka 2018).

Despite the magnitude of the consumption savings, a recurring finding is the high heterogeneity in the impact estimates. Apart from differences in experimental setups, both socio-demographic and psychological variables can explain the variety in consumption levels and response to information feedback (Abrahamse and Steg 2009). Nonetheless, often the relationship between these variables and treatment effects is not evident (Houde et al. 2013), and larger studies are still required (Andor et al.2022).

2.2.3 Interventions for specific end-uses

Besides nudges for achieving energy conservation, some programs specifically focus on reducing heating or cooling. In Myers and Souza (2020) authors study the effect of HERs repeatedly mailed to household with nudge for reducing heating. The experiment is introduced in a college resident, where tenants do not pay for the energy bills. Authors find almost no behavioral changes for heating demand, suggesting that behavioral interventions may not be sufficient in the absence of monetary incentives. On the contrary, despite little monetary incentives, an intervention to reduce the indoor temperature in a Swiss canton was successful (Kandul, Lang, and Lanz 2020): even though building-level heating costs are shared across flats, the field experiment estimates a reduction of average indoor temperature of $-0.28°C$. In another field experiment conducted in a graduate residence at the National University of Singapore (Brülisauer et al. 2020) authors provide appliance-specific feedback on electricity consumption from air-conditioning usage only, achieving a reduced electricity consumption from air-conditioning of 17%.

Other types of behavioral intervention (Burkhardt, Gillingham, and Kopalle 2019) focus on encouraging energy conservation during peak load days (in particular reducing air conditioning on hot summer days) while electricity cost is reduced when there is abundant renewable generation, to foster electric vehicles load in these times. The purchase of green energy is another action that should be encouraged; presenting green energy as default option, for both business and private sectors, leads to an 80% of customers to keep the green option, in a stable way for at least 4 years (Liebe, Gewinner, and Diekmann 2021).

2.3 Machine learning for causal inference

The literature offers well-established statistical and econometric approaches and tools to determine the causal link between an intervention and its effect on some outcome of interest (Angrist and Pischke 2008). At the core of the problem lies the prediction of a counterfactual, i.e., what would have happened if the individuals treated with an intervention had not been treated. The control group of a well-designed randomized experiment offers a counterfactual of the highest quality, as potential confounding factors cancel out when comparing treated and non-treated groups. Unfortunately, a perfectly randomized experiment at the desired scale is not always socially, technically or economically feasible.

Several tools have been developed to deal with natural or other types of experiments that are not perfectly randomized. These include linear regressions, instrumental variables, regression discontinuities, and differences-in-differences (Varian 2016).

Machine learning (ML) is a field of study of computer algorithms that learn through experience without being explicitly programmed (Samuel 1967). Different categories of ML can be used depending on the problem, such as classification, regression, clustering, and reinforcement problem (Alzubi et al. 2018). ML enters the picture of causal inference as an additional tool to compute the counterfactual needed to estimate the causal effect of an intervention. Its benefits are particularly relevant with large amount of data, like in the case of smart meters readings. First, traditional methods might not be as computationally efficient as ML techniques to manipulate big data. Second, machine learning techniques offer more flexible models to capture non-linear relationships than traditional linear models. Third, ML can be useful in identifying features that are more relevant among the possibly many available (Varian 2016).

Causal inference is one of the most promising areas of collaboration between ML and econometrics (Varian 2014; Mullainathan and Spiess 2017). Two aspects where ML can offer improvements over traditional approaches are predictive modeling and model uncertainty quantification. First, a good prediction of a counterfactual allows for a better estimation of the average treatment effect, and most ML tools are focused on optimizing prediction accuracy. Second, ML is associated with running a large number of alternative specifications, it helps to separate the data ensemble in meaningful subspaces with peculiar characteristics and relationships, and it often leverages on non-parametric methods for quantifying uncertainties (e.g. via bootstrapping). ML can thus offer valid alternatives to characterize uncertainties and explore heterogeneities of causal inferences. Nonetheless, we should remember that no matter how sophisticated the causal estimation strategy is, eventually a randomized controlled trial remains the first-best way to validate its accuracy.

In the following sections, we present several potential methods and applications of ML to the problem of estimating the impact on energy conservation of behavioral interventions. Eventually, causal inference is conducive to empirically based policy making, and machine learning can help to optimize such data-driven decision-making (Athey 2017).

3. Novel impact evaluation methods

3.1 Counterfactual forecasting

In settings where for some reasons it is not possible to perform randomized control trials, as with observational studies, it is crucial to find alternative ways to deduce counterfactuals for the individual treatments. Many studies overcome the problems of a missing control group with counterfactual prediction, often obtained with machine learning approaches (Abrell, Kosch, and Rausch 2021). Comparing results obtained from randomized experiments for electricity consumers with those obtained with prediction algorithms, counterfactual forecasting replicates the treatment effects obtained with RCTs (Brian C. Prest and Palmer 2021; Burkhardt, Gillingham, and Kopalle 2019).

In the context of a natural experiment, like the first COVID lockdown, forecasting is the only way to build a counterfactual scenario to compare with; for example, in Granella et al. (2021) machine learning tools have been used to estimate air pollution reduction during the lockdown by comparing the time series of concentration of $PM_{2.4}$ and NO_2 with the values obtained by the model (that had been trained on historical data, from 2012 to 2019).

In Burlig et al. (2020) authors estimate energy savings obtained with energy efficiency upgrades in K-12 schools in California. They estimate with both a panel fixed effects approach and a machine learning one, with the second giving more accurate results. They finally studied the response heterogeneity but they could not identify school characteristics that systematically predict a better response.

Machine learning algorithms can accurately predict counterfactuals, leading to an estimation of treatment effect heterogeneity (Section 3.4) also in studies with staggered treatments. In Souza (2019), the author underlines that rich data availability allows for an accurate prediction of counterfactual, leading to a good recovering of treatment effect, and he successfully applied the approach to evaluate an energy efficiency program in the US.

3.1.1 Power load forecasting

To study the effect of behavioral intervention using smart meter data, the problem of computing counterfactual is related to that of electric load forecasting. The forecasting of electricity consumption is an important problem, very relevant for industries and academia. It is relevant at different scales, from hourly forecasting to long-term, aggregate consumption (Hong and Fan 2016).

Hourly electricity demand at an aggregate level presents recurrent patterns at a daily, weekly, and yearly frequency. Methods based on Fourier analysis achieve good results in predicting both short-term (hour or day ahead) and long-term (year-ahead) predictions (Yukseltan, Yucekaya, and Bilge 2020). It is important to have at least a two-year observation period to make long-term prediction. It is also fundamental to set aside a time interval for validation of prediction.

Prophet is a procedure for time series forecasting developed by Facebook, with available open-source implementations in Python and R. Time series are forecast with an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. The algorithm can automatically detect trends and change point in the data; season components are modeled with Fourier series. In scientific literature Prophet is widely applied to forecast time series in different fields; for electricity market prediction, it is suitable to predict short- or long-term aggregate load demand of a city, e.g. Chicago (Parizad and Hatziadoniu 2021), or a country, e.g. Kuwait (Almazrouee, Almeshal, Almutairi, Alenezi, Alhajeri, and Alshammari 2020; Almazrouee, Almeshal, Almutairi, Alenezi, and Alhajeri 2020). Prophet is used also for the forecast of buildings consumption, like shopping malls and office buildings (Gong et al. 2020).

While prediction of load demand at aggregate levels achieves good results when enough historical data is available, the prediction of single tenants' consumption is subject to a greater variability in the time series as well as in the forecast accuracy (Shapi, Ramli, and Awalin 2021). Merging residential buildings together and comparing different methods is recommended: in Panigrahi (2020) the author compared the performance of Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX), Seasonal Autoregressive Integrated Moving Average with Explanatory Variable (SARIMAX), PROPHET and Long-Short Term Memory (LSTM) for forecasting average energy demand of a group of residential buildings in London, finding that LSTM performs best. Comparing algorithms is nevertheless not so trivial, and new approaches to load forecasting keep being applied. In general, the granularity and the time-horizon influence the effectiveness of the forecast: in Oreshkin et al. (2019) the N-beats neural network has been used to forecast mid-term electricity load (with monthly granularity and a forecast horizon of 12 months). One week of learning data to forecast one day is achieved with Temporal Fusion Transformers in Lim et al. (2021). Finally, different models should be merged together to achieve even better accuracy via models ensembles (C. Li et al. 2019).

Machine learning algorithms that were not designed for time series can be extended to time series forecasting. XGBoost (T. Chen and Guestrin 2016) is an algorithm which predicts the future values of a time series based on selected temporal features extracted from the data. Despite that the algorithm was not designed for time series, it can be successfully applied to temporal forecasting whenever the temporal features extracted from the data are meaningful for predicting the signal. On one hand, the main limitation of XGBoost is the difficulty in predicting trends. On the other hand, it offers the advantage of knowing which features are important for prediction. For example, the weather is increasingly impacting electricity demand and is therefore an import predictor for power load (Staffell and Pfenninger 2018).

In summary, to perform time series forecasting it is important to properly preprocess data (depending on the model adopted, normalization may not be necessary), divide the data in training and validation part (historical forecast are useful to assess the validity of the model), include relevant covariates when available (e.g. the weather), and apply different methods to check the best performance in the specific dataset, eventually applying more than one model together. The model could be as simple as a linear regression, to the classical ARIMA, the Facebook Prophet algorithms, or more advanced neural network models. For all of these functionalities open-source implementations can be found, like DARTS in Python (Herzen et al. 2021).

3.2 Clustering

Availability of hourly consumption data can provide further insights on how people behave. Machine learning, and in particular clustering, has been applied in the past to better understand patterns of consumption. Although several approaches are documented (Y. Wang et al. 2015), some are of more immediate application and interpretation, like the one by Kwac, Flora, and Rajagopal (2014), focusing on the direct clustering of normalized load shapes.

Clustering have mostly been used in the context of static load profiling rather than for detecting behavioral changes. More recently, clustering has been applied also for detecting behavioral changes. For example, clustering was used to assess the impact of natural experiments such as the COVID19 pandemic and subsequent lockdown (García et al. 2021, Abdeen et al. 2021).

Daily electricity consumption shows high variability between and within households, for different days of the week and throughout the year. Being able to identify a few exemplary daily behaviors, which are not too dissimilar to the multitude of possible profiles, allows to find changes in consumption patterns. Clustering can thus be used for assessing the impact of behavioral intervention in the consumption habits, as a change in daily load shapes. In fact, changing part of the consumption load to off-peak periods brings both economic and environmental benefits, particularly in the context of economies which rely more and more on intermittent renewable energy provisions.

One of the most popular clustering techniques is k-means (Lloyd 1982), which divides data into clusters characterized by their typical shapes called centroids. The algorithm minimizes *inertia*, which is obtained by summing the squared distance between each data point and its closest centroid. The number of clusters needs to be defined a priori, and the most appropriate one can be chosen by comparing the final inertia through the elbow method (Thorndike 1953). K-means is suitable for load profiles clustering, because it is relatively simple, it typically finds meaningful centroids profiles, and it scales well with the amount of data. K-means is the chosen clustering method in this report (see Section 4-6). We refer the reader to newTRENDs deliverable 5.1 (Marangoni et al. 2022) for a detailed report on clustering of residential load profiles.

3.3 Explainable machine learning

Machine learning (ML) algorithms and in general artificial intelligence is increasingly being adopted in a wide range of sectors, from industry to academia, achieving better and better performance over time. This widespread success is due to a combination of more powerful machines and easier access to a vast amount of data. The complexity is also increasing though, and often ML approaches are considered as ’black box’, leading to difficulties in the interpretation of results. The lack of interpretation may reduce trust in results. To overcome this drawback, the field of Explainable Artificial Intelligence (XAI) develops methods to explain and interpret machine learning models (Linardatos, Papastefanopoulos, and Kotsiantis 2021). In this context, the term explainability and interpretability are used interchangeably, but they actually have a different meaning: interpretability is defined by Miller (Miller 2019) as "the degree to which a human can understand the cause of a decision". On the contrary, explainability regards the understanding by humans of the internal procedure of the machine learning model in giving the output.

The SHAP (SHapley Additive exPlanation) was introduced by Lundberg and Lee (Lundberg and Lee 2017) to explain the output of any machine learning model. The method is inspired by game-theory and it computes the importance value that the machine learning model places on each feature for predicting each data point; positive or negative SHAP values indicate the direction of the effect (i.e. if the feature influences the predicted value by increasing or decreasing it). It therefore guarantees both global interpretability (it shows how much each predictor contributes) and local one (for each observation it is possible to derive its set of SHAP values), and has the properties of local accuracy, missingness, and consistency (Lundberg and Lee 2017).

SHAP, together with LIME (Lundberg and Lee 2017), is one of the most dominant methods in the literature to assess feature importance, it is model-agnostic and can be applied to any type of data. For example, in R. Li et al. (2020) SHAP values unravel nonlinear relations between mortality risk and individual risk factors, for the prediction of mortality risk in prostate cancer; in D. Wang et al. (2022) XGBoost proves to be successful together with SHAP in investigating how to improve process management in wastewater treatment plants; in Akhlaghi et al. (2021) SHAP method is used to interpret the contribution of the operating conditions on performance parameters of Guideless Irregular Dew Point Cooler. However, there is not an established way to interpret statistical significance of SHAP values. Furthermore, similarly to any ML methods, the complexity of the algorithms does not automatically overcome selection bias and variation in data collection methods.

XAI has also been applied to electricity consumption time series data by providing visualization of highly personalized electricity consumption feedback, obtained with SHAP and LIME (Wastensteiner et al. 2021). In an experiment with 152 participants, results show that humans can assimilate the patterns displayed by XAI visualizations, but standard visualizations can make the feedback better understandable by users.

XAI is an interesting tool for assessing behavioral interventions on smart meter data. In fact, electricity consumption can be modeled and predicted with an algorithm such as XGBoost (see Section 3.1.1), and SHAP values help unraveling the importance of the features used to predict consumption. In this way, it would be possible to estimate the influence of e.g. the weather, or receiving a home energy report, on the power consumption. We will illustrate an application of this approach in Section 4.2.2.

3.4 Heterogeneous treatment effects

How does a treatment effect vary across individuals? In many studies behavioral interventions appear to be more effective for a subset of the population (i.e., nudge effectiveness on immunization campaign in India (Banerjee et al. 2021; Chernozhukov et al. 2018)). Understanding which subgroup responds better to a treatment allows for improved targeting and design of treatment. For example, a randomized control trial (RCT) experiment in the USA showed how political liberals respond to HERs (Home Energy Reports) by reducing consumption two to four times more than conservatives (Costa and Kahn 2013). In Asensio and Delmas (2015) authors find that residential households that receive tailored information about environmental and health damages produce 8% energy savings compare to the control group. For family with children instead, the average effects were found to be 19%.

Estimating the average treatment effect (ATE), which is the common approach with randomized control trials (RTCs), gives only limited information, ignoring the variability in the response to treatment. What is in general more interesting is the conditional average treatment effect (CATE), which tells the variation in the response for different subgroups. This variability allows to personalize treatment and better understand causality mechanisms. Traditional econometrics provides various methods for estimating the CATE. A simple approach is to compute the ATE among subjects in a specific subgroup. Alternatively, a regression framework together with dummy variables differentiates the effects on subgroups. In cases in which multiple subgroups may be tested to check for differences in the treatment effects, we encounter the problem of multiple comparisons: statistical significance become more likely to be achieved, leading to an increase in probability of wrong results. To mitigate the problem, it is possible to either reduce the number of subgroups tested (or pre-specify the tests that you will carry on through a registered pre-analysis plan: PAP), or to adjust the p-value to account for the simultaneous test on multiple hypotheses (e.g. with Bonferroni correction, in this case the risks is to be too conservative since the hypotheses are considered independent).

With the recent availability of big data, the number of covariates and interaction terms can outnumber the observations, and it may be interesting to span all the possible subgroups to check for heterogeneity, without relying on a-priori assumptions. In this context, machine learning methods have been developed to automate the search for heterogeneity in treatment and allow for cross-validation of the outcomes (Athey and Imbens 2015). For example, metaalgorithms (or meta-learners) build on basic algorithms, like logistic regressions, random forests, XGBoost …, to estimate the CATE. The most popular meta-learners are S-learner, T-learner, X-learner (Künzel et al. 2019), with Python implementations available online (H. Chen et al. 2020).

3.4.1 Causal forest

A generalized random forest (causal forest) is a causal machine learning method developed by economists Susan Athey and Stefan Wager (Athey, Tibshirani, and Wager 2019). Causal forests are flexible nonlinear models able to evaluate heterogeneous treatment effect, applicable also with a high number of features and able to provide confidence intervals. The algorithm consists of recursively partition a sample in subgroups that maximize heterogeneity across splits, with optimization of treatment effect heterogeneity as splitting criterion. Each run of the partition is a classification and regression tree (CART), while the forest is constituted by multiple trees which analyze a bootstrapped subsample of data. The algorithm of causal forest is an adaptation of random forests to the study of causal effect. While in a random forest the same dataset is used to create the tree structure and evaluate the ATE, Athey and Imbens (2015) introduce the concept of *honest estimation*, where each bootstrapped subsample is split in one part used to create the tree, and the other to estimate the treatment effect for each leaf. Causal forests have been used to estimate heterogeneous treatment effects in different fields, such as road safety (Zhang, Li, and Ren 2022). A Python implementation of the causal forest algorithm is available in the *econml* package from Microsoft (Battocchi et al. 2019).

Causal forests have been applied to evaluate target treatment for household energy use in Knittel and Stolper (2019), where across fifteen Opower waves the ATE corresponds to a reduction in monthly electricity usage of 1 percent (9 kWh). Data consist of one year of pre-treatment and 3 years of treatment data with monthly resolution. In the first two months the authors find no significant impact, but from subsequent months a consistent downward trend in consumption is observed. Causal forests have been applied to investigate which features lead to higher energy savings, finding that the stronger predictor is pre-treatment consumption.

A RCT with 120’000 customers in Germany found conservation effects of information campaigns for residential energy conservation (Andor et al.2022) ranging from zero to −1.4%. Using causal forests, they show that heterogeneity across utilities cannot be explained by socio-demographic characteristics.

3.5 Conclusions

In the previous sections we described multiple innovative techniques based on ML that can greatly improve the assessment of behavioral interventions. The techniques to apply depend on data availability and research goals. For example, if a control group is not available, but the pre-treatment period is long enough compared to the natural periodicity of the signal (typically one year for power consumption), counterfactuals can be constructed by forecasting the future of the pre-treatment periods (Section 3.1). Domain-specific knowledge is necessary for choosing the best forecasting approach in the context of power load data (Section 3.1.1).

Clustering techniques, despite being usually applied for descriptive purposes, are also helpful to underline effects of an intervention, e.g. to test if the intervention changes consumption patterns differently for multiple groups of people (Section 3.2). While clustering of load curves is helpful to describe patterns of behaviors, and changes thereof, other ML approaches specifically aim at untangling heterogeneity (Section 3.4). In particular, causal forests (Section 3.4.1) are suitable for exploring the heterogeneity of the treatment effect on a high number of variables. Finally, Explainable ML developed indicators that are able to assess features importance for a ML model. This approach is very promising for evaluating treatment effects, while more work is needed to accurately establish findings' significance.

In general, ML techniques are advantageous for many reasons. ML models are more sophisticated, with linear and nonlinear interactions taken into account in an automated way, as well as high-order interactions. This gives better forecasting, with a high out-of-sample predictive power. The assessment of heterogeneous treatment effect is more accurate with ML approaches like causal forests and ML algorithms usually are suitable for studying very large datasets (Athey and Imbens 2019). Using clustering techniques, it is possible to detect exemplary behaviors on the data.

More theoretical work is still needed to develop methodologies to estimate robust standard errors, comparable to well-established econometric standard approaches (Mullainathan and Spiess 2017). Moreover, data need to be comprehensive enough to allow application of ML tools like forecasting (the series of historical data needs to be long enough to give accurate predictions); while for assessment of heterogeneity, the relevant variables need to be available, and for each group of users there needs to be sufficient representativeness.

In the remaining part of this report we show three case studies where we apply clustering techniques in different ways. We also show an application of explainable ML methodology for assessing HER treatment effect (Section 4.2.2).

4. Italian case study 1 - Bologna

4.1 Impact of COVID-19 restrictions on household consumption patterns

As discussed in Section 1.2, COVID-19 lockdown had a strong impact in electricity consumption, both in the domestic and industrial sectors. In particular, the increase in working-from-home, induced by the pandemic lockdown, has reinforced the societal trend of increase in home-office already happening in modern societies. For this reason, the first COVID-19 lockdown is a unique natural experiment that allows us to foresee the consequences of a high share of people working from home. For the scope of NewTrends, exploring the influence of lockdown on energy demand is related in particular to Task 6.2 (Modeling the impact of digitalization and new market trends in the tertiary sector on energy demand and energy-efficiency potentials).

In this section we evaluate the impact of COVID19 on household power consumption in Italy, using smart meter data. The dataset is composed of hourly smart meter data recordings from thousands of households from the area around the city of Bologna in Italy. Recordings are available for the year 2019 and 2020, although the households for the two years are not the same (only two of them are in common) and for 2019 we have a lower number of users. We use a clustering approach (*k*-means), to be able to find groups of users that behave similarly, thus unraveling the difference in hourly and daily consumption before, during and after the first national Italian lockdown.

Our aim is to evaluate the impact of the multiple closures of activities and lockdowns on the average daily consumption of domestic users. Since electricity consumption changes depending on seasonality, day of the week, and special holidays (e.g., Easter), it is crucial to make a comparison between the consumption in 2020 and the trend of the previous year. Although in our dataset the sample of households in each year is different, we are still able to get interesting insight from the comparison.

The preprocessing of data consists of the following steps. We eliminate duplicate data, discard data of zero consumption, and days with less than 24 hours of valid data. Finally, we remove users who have more than 5 days of missing data in the time interval under study.

We select the time period between the first of February and the end of May, for both 2019 and 2020. For each household, we use the time-series of the (z-standardize) average (of the log of) daily consumption. In order words, we get a signal with the average of electricity consumption for each day in the four months. For 2019, in the period between the first of February and the end of May, we have 119 valid days; the users are 165. For 2020, in the same period we have 120 valid days, and 518 households. We select only household in the city of Bologna.

We divide the period February-May 2020 in four time intervals, and we project the same intervals on 2019 for comparison:

- *Before* COVID-19: From the first of February to the 9th of March;
- *First* part of Lockdown: until the 29th of March. We choose this date because it is the daylight-saving date for 2020; in this way we can map difference in behavior due to seasonality.
- *Second* part of Lockdown: until the 4th of May, when reopening of activities started in 2020.
- *After*: reopening, until the end of May. It goes from the reopening of activities, until the end of the month.

For the identification of consumption patterns, we cluster the signals, using k-means algorithm (Lloyd 1982), and the elbow method (Thorndike 1953) to determine the number of clusters. Throughout the analysis we consider two time scales of signals to cluster: the *daily consumption for the whole period February-May* (2019 and 2020), and the *hourly consumption for a day* - selected for the same period. While the whole-period-clusters shed light into the overall variation in consumption throughout the pandemic phases, the day-clusters are helpful to unravel daily habits of people.

First, we discuss the analysis of the day-clusters: hourly consumption for a day. In Figure 1 we show the clusters of daily consumption, at hourly resolution, obtained with k-means clustering. The daily clusters belong to the period of February-March of either 2019 or 2020. The daily profiles for different users of both years are merged in one dataset in order to find common daily behaviors. Typical daily consumption patterns are in fact common in the two subsequent years, as one would expect; only the frequency of each type of daily behavior changes, in particular as a consequence of the lockdown, as we will describe late. The number of clusters, six, is selected trough the elbow method. In Figure 2 the distribution of daily energy consumption of the curve in each cluster is shown. Both these plots help understanding how the clusters describe hourly behaviors of household.

- 1: Afternoon-evening consumption;
- 2: Small morning peak, high (early) evening consumption;
- 3: Late morning - early evening activities;
- 4: Uniform, low consumption;
- 5: Late evening activity;
- 6: High lunch peak - medium dinner consumption.

In Figure 3 we illustrate the fraction of daily curves which belong to each cluster, for each of four corresponding time periods in 2019 and 2020. From the change in frequency distribution of clusters we see how daily consumption 2 (small morning peak, high early evening consumption) and 5 (late evening activity) decrease during lockdown: those behaviors characterized a common 8-hours working days - outside home. Early evening consumption (2) remains of low prevalence also in the reopening period, while the evening consumption (5) regains comparable level with the previous year (towards the summer, daily consumption (5) was monotonically increasing also in 2019). The frequency of cluster (3) (intense morning and evening consumption) does not change significantly, while cluster (4) (uniform and low consumption) shows a slight decrease in the second part of the lockdown, probably indicating that people do not go on holiday). Finally, clusters (1) and (6) (Figure 1) increase their prevalence during lockdown, and even in the reopening time these clusters remain more prevalent than before. In summary, daily energy load (1) and (6) (high afternoon-evening consumption and intense lunch-time load) are likely load curves that characterize smart working.

Figure 1: Cluster of daily consumption at hourly resolution. Data cover the period of the first COVID lockdown (from February to May) in 2020, and the same months of the previous year 2019. In the squared brackets, the timing of the peaks.

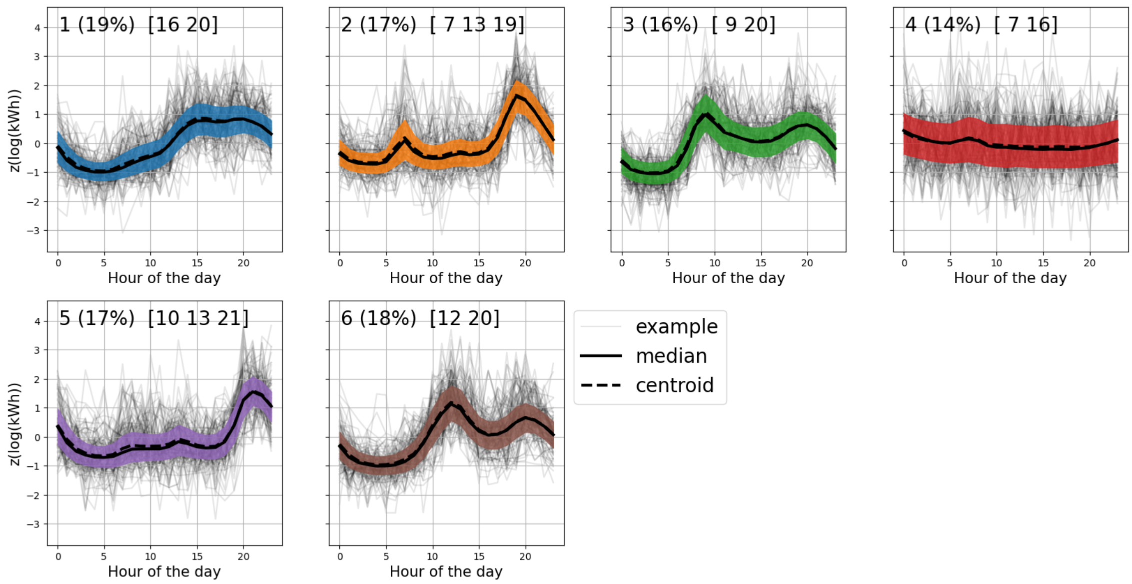


Figure 2: Distribution of daily total energy of the curves in each cluster. The bars span the range of values.

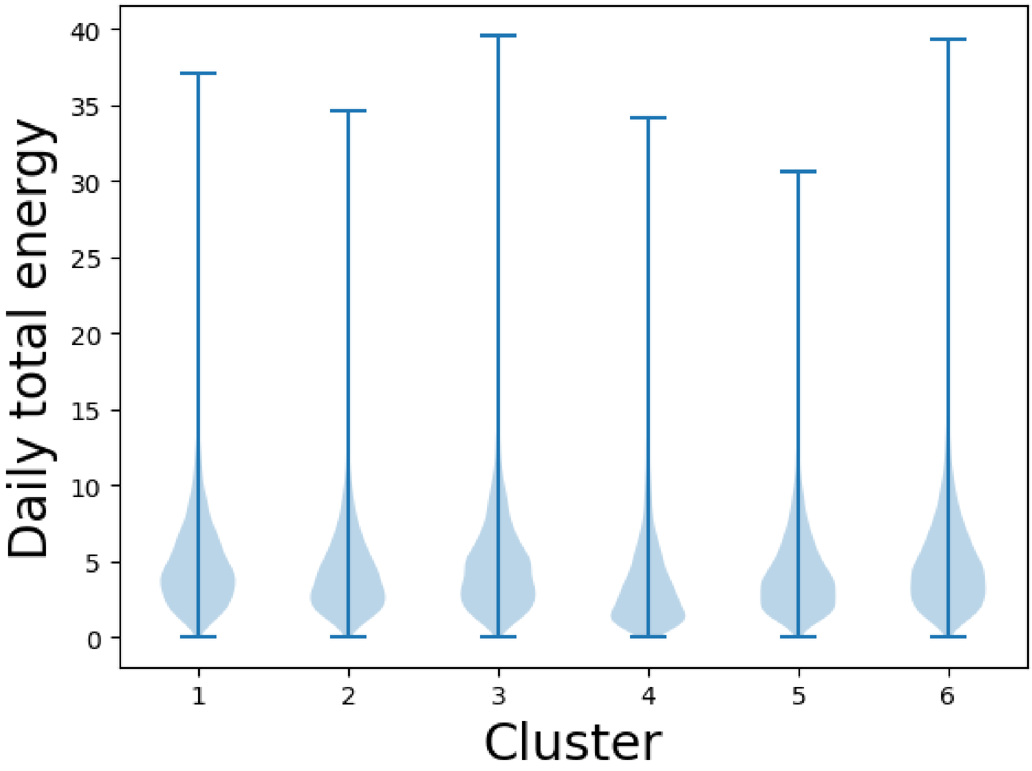
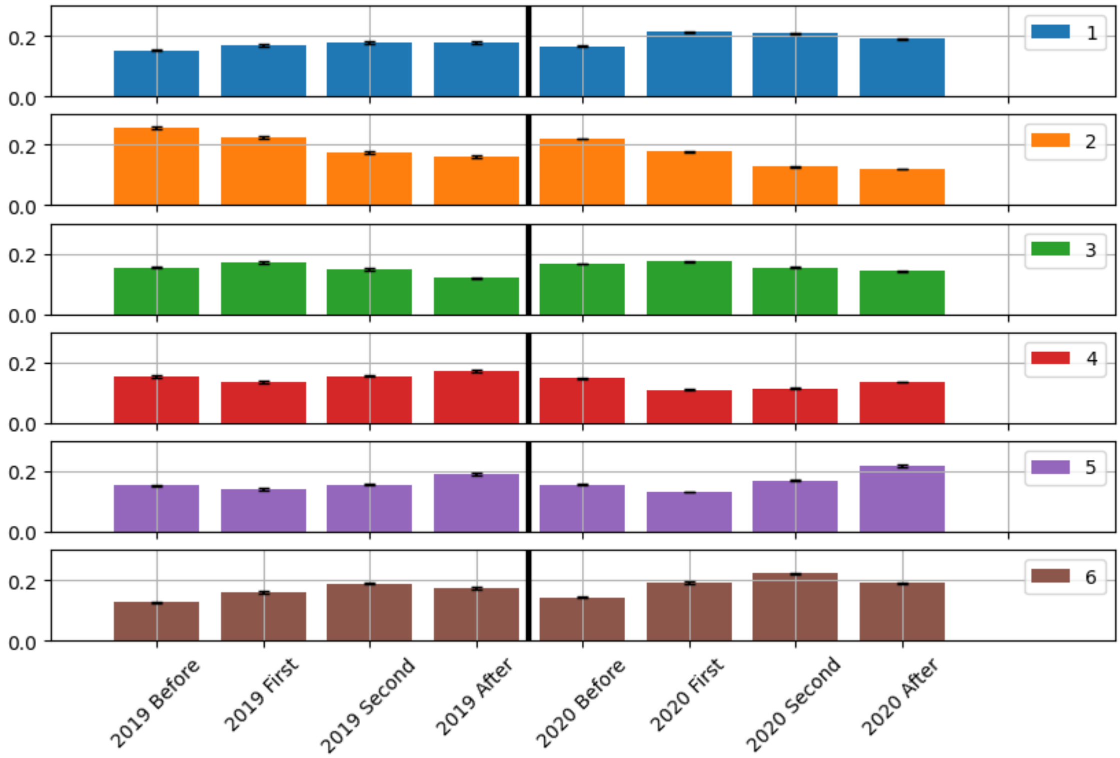


Figure 3: Frequency of daily curves belonging to each cluster, for four corresponding time periods in 2019 and 2020. The black line on top of each bar indicates the -almost invisible- standard deviation obtained with bootstrap.



With day-cluster hourly consumption analysis we show the average variation in daily behaviors. It is also interesting to characterize households depending on how the overall consumption changes with the pandemic, and eventually link these changes with possible changes in daily consumption at hourly resolution. We analyze these results for the daily consumption clusters separately for 2019 and 2020. For the 2020 we find an optimal value of 6 clusters, while for the 2019 we keep the same value for comparison (the optimal value is 7). Figure 4 and Figure 7 show the clusters obtained for the two datasets. We underline the date correspondent to daylight saving and Easter. For the 2020, the year in which lockdown and restrictions started to take place in Italy, we underline also the time interval when there was the first COVID-19 national lockdown (09/03/2020 - 04/05/2020). The 9th of March 2020 was when the national lockdown was declared, with the closure of schools, activities and shops (a part from grocery stores, pharmacies and other essential services), and the citizens were obliged to work from home (with the exception of *essential* workers); while on the 4th of May, many activities started to open again after the lockdown. In Table 2 we provide some information about house characteristics. Unfortunately, this information is available only for 2020. We show the percentage of Green electricity contracts, the percentage of households where tenants are resident, and the average surface of the house in square meters.

Table 2: Percentage of household: with a Green contract and with Resident occupants; average dwelling Surface in square meters; these values are the average for every cluster of Figure 7. The average values for all the sample are also shown.

Cluster	Green	Residents	Surface
average	37%	88%	78.39
a	33%	86%	79.27
b	27%	89%	78.73
c	43%	90%	72.76
d	30%	86%	77.70
e	33%	91%	88.48
f	38%	90%	77.27

From the cluster representation, we can see clearly that the Easter valley (of cluster (a) and (e) in 2019, Figure 4) is absent in all the clusters for 2020 (Figure 7), indicating that obviously people in 2020 have not been able to go on vacation on those days. For both the years we have a cluster (cluster (d) for 2019, (a) for 2020) in which the difference between weekday and weekend is particularly evident and regular throughout the period. This cluster indicates people likely to not be as much at home during the week as during the weekend; the same is probably valid also for the households in the cluster (a) for 2020 - despite the restrictions, some people where still going to work outside home. For 2020 there is a peak around the beginning of the lockdown (cluster (b) and (d) in Figure 7, synchronized with the Daylight-saving date) followed by a decrease of consumption. Cluster (c) indicates probably household with people doing smart working, since the consumption has increased regularly with the first closures and then has continued to stay high. The information in Table 2 suggests a smaller dwelling surface than average and high prevalence of green contracts, likely to indicate relatively recent electricity contracts, in a central area of the city. Cluster (e) has a lower level of consumption during the closure, and it is also the cluster with the minimum mean household consumption (Figure 8); it likely indicates houses where some tenants moved somewhere else during the lockdown, for example off-campus students. In fact, looking at Table 2, households in cluster e are likely to be bigger than average. Overall, the trends evident from the 2020 daily clusters (Figure 7), are quite different from the clusters of 2019 (Figure 4). For the 2019, apart from Easter and the week-weekend regularity, clusters just show some increasing or decreasing trends, probably related to seasonality through heating and cooling systems. For the 2020 instead, daily clusters allow to characterize groups of households based on their response to closure.

Figure 4: Clusters obtained with k-means for the daily consumption over a period of a few months, for the year 2019.

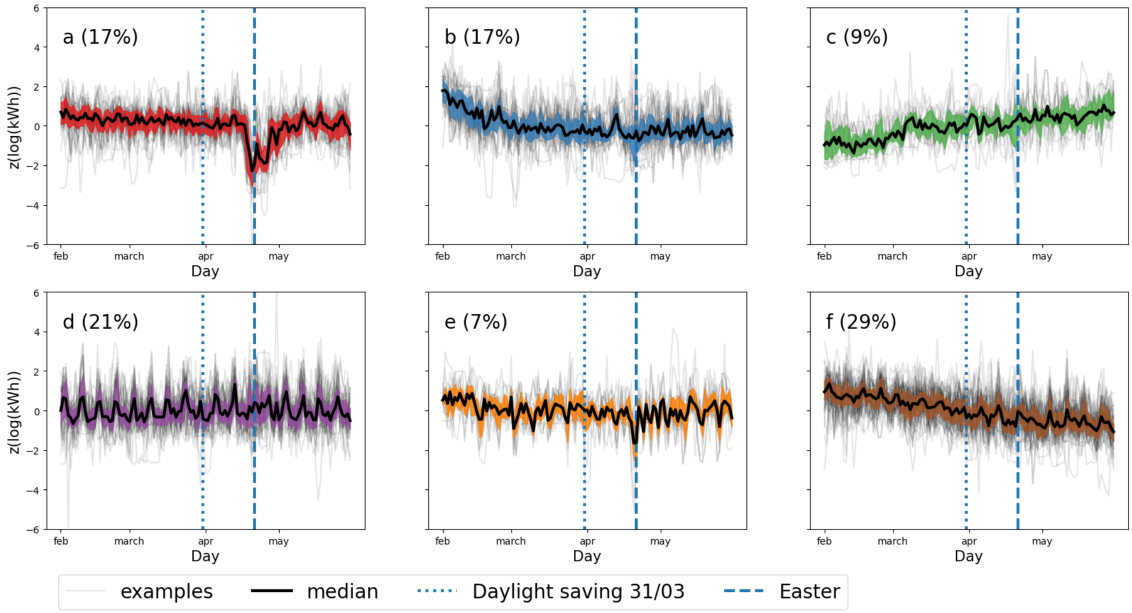


Figure 5: Distribution of mean household consumption for consumers of each cluster in Figure 4. The bars span the range of values.

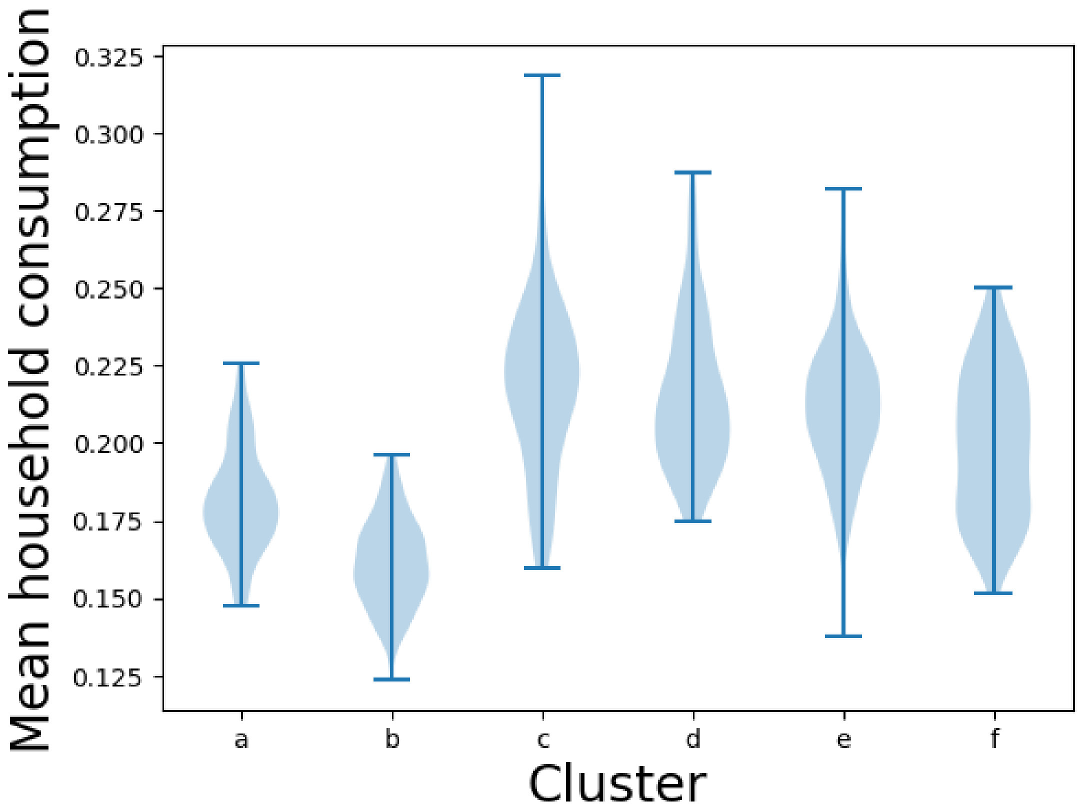


Figure 6: For each daily cluster in Figure 4, the distribution of hourly cluster (from Figure 1). Errors bars represent standard deviations obtained with a bootstrap procedure (10 resample, each with 60% of users). Gray bars represent frequencies of each cluster.

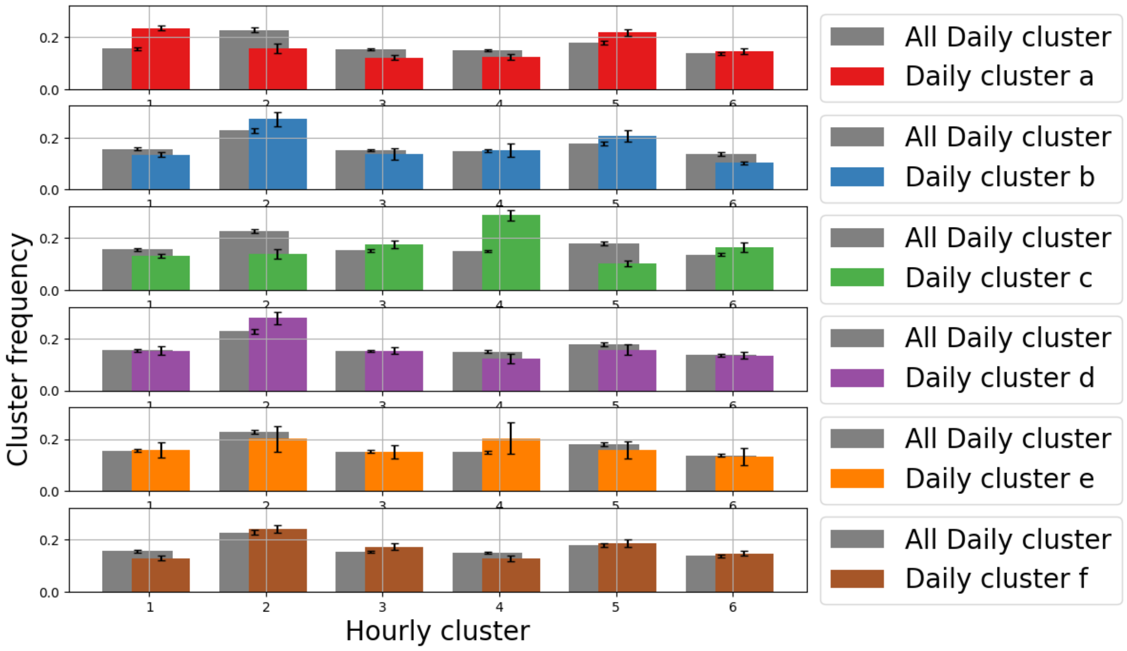


Figure 7: Clusters obtained with k-means for the daily consumption over a period of a few months, for the year 2020.

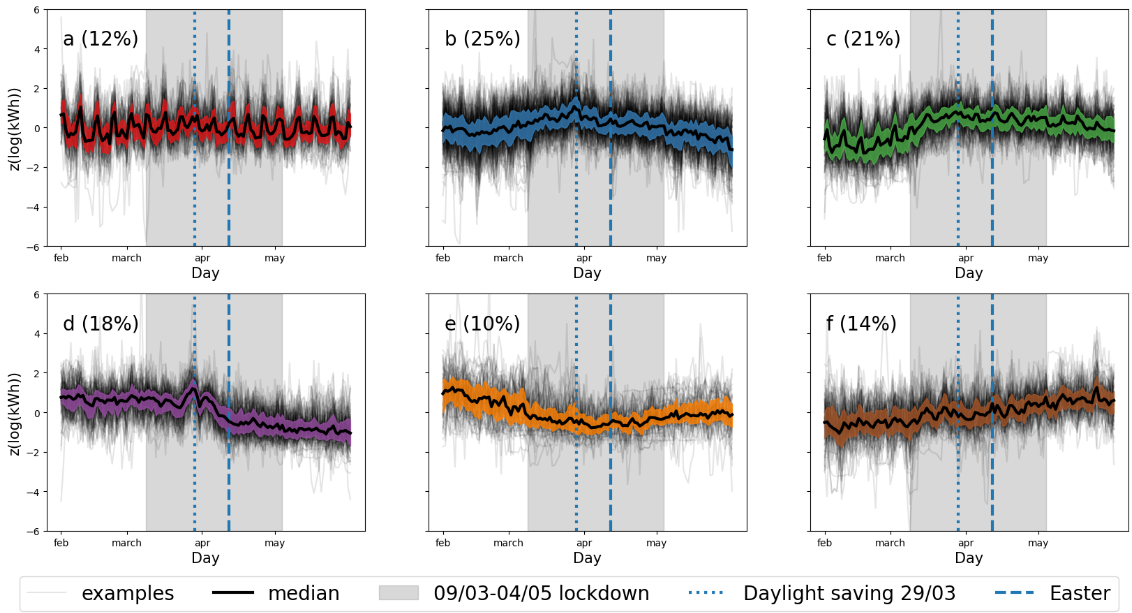


Figure 8: Distribution of mean household consumption for consumers of each cluster in Figure 7. The bars span the range of values.

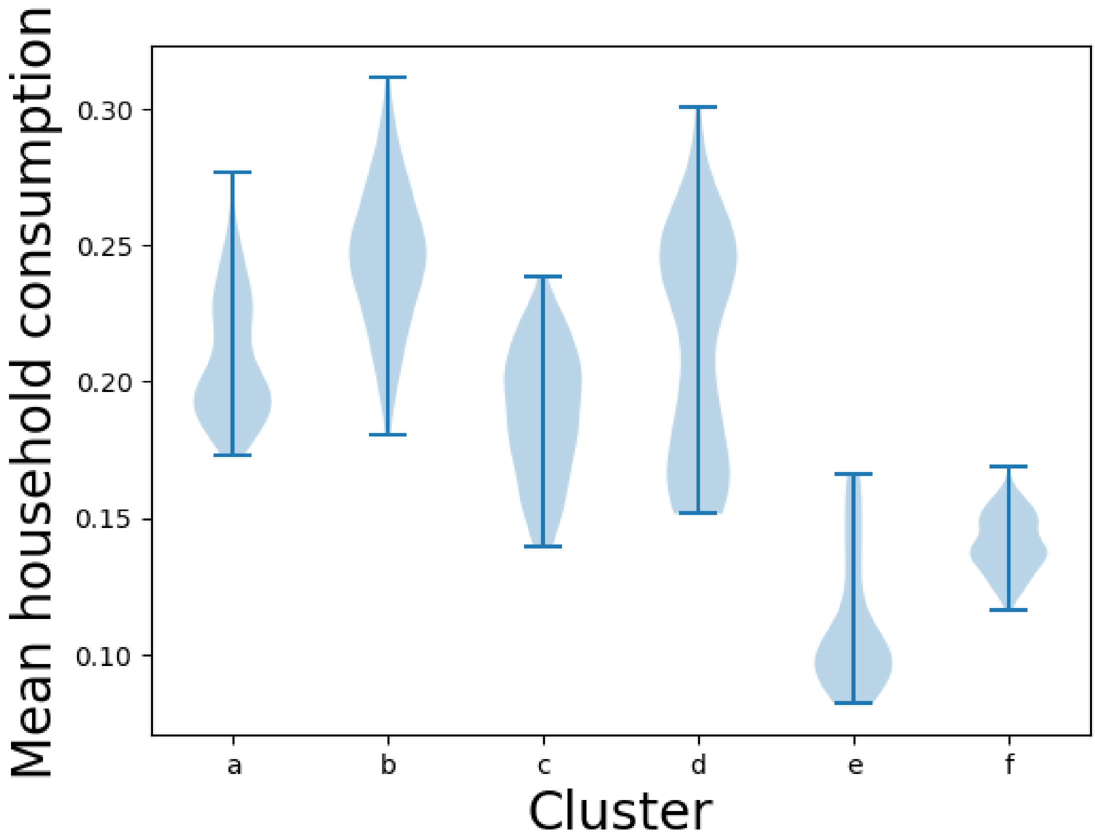
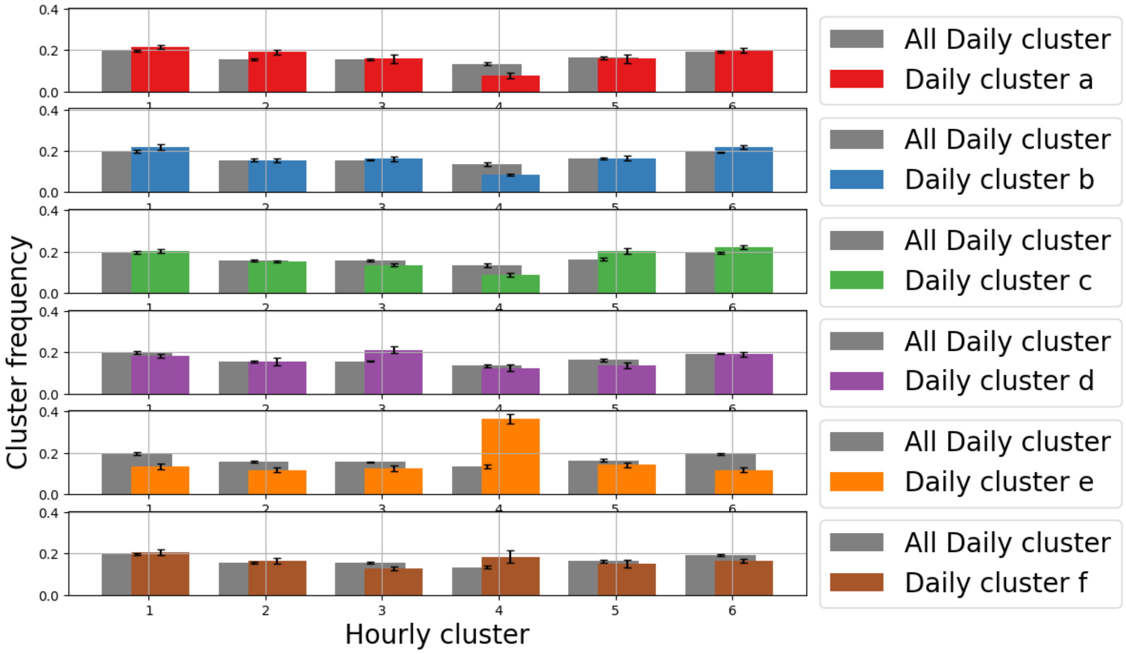


Figure 9: For each daily cluster in Fig. 7, the distribution of hourly cluster (from Figure 1). Errors bars represent standard deviations obtained with a bootstrap procedure (10 resample, each with 60% of users). Gray bars represent frequencies of each cluster.



Finally, we associate daily and hourly clusters, to link the monthly behaviors with the daily patterns: we check for each households' types a-f (years 2019 and 2020) if some daily clusters at hourly resolution are more prevalent than others; this analysis indicates different typical profiles of users throughout the pandemic lockdown. In Figure 6 and Figure 9 we show the different frequencies of hourly cluster 1-6 (from Figure 1) for each daily cluster a-f. Comparing the bars with the gray ones, it is possible to check if the specific hourly cluster n (with n between 1 and 6) is more/less common on daily cluster m (with m between a and f) compared to the average frequency on all the possible daily clusters. We highlight here some features that appear from those comparisons. For the 2019, we underline how the regular week-day/week-end behavior of household type (d) is associated with a prevalence of hourly cluster (2) (Figure 1), the typical eight-hours-day-at-work behavior. For the year 2020, we notice how household type (c) (forced smart-working users) have a high prevalence of hourly behavior (5) and (6) (from Figure 1), describing late morning peak (behavior 5) and a high peak at lunch-time (behavior 6). It is reasonable to expect the high peak at lunch-time hourly behavior as more typical for people working from home.

4.1.1 Conclusions

In the previous section we applied different clustering techniques to associate daily and hourly clusters of residential consumption throughout the period before, during and after the first COVID-19 lockdown in Italy. Broadly speaking, we were able to identify households that keep a similar regular pattern of consumption as before lockdown, and others who switch to a working-from-home consumption pattern, with a strong increase in mid-day consumption peak.

The method proposed of linking clusters obtained with multiple time periods and resolution is a valuable approach that can be generalized for others similar analysis of natural experiment when a control group is not available. In fact, it makes it possible to cluster users at a time scale of the experiment (in our case, daily consumption spanning the period from before to after the event) thus recognizing the groups of people who respond in different ways to the natural event. Separately, consumption can be clustered at a smaller scale (e.g., consumption of a day at hourly resolution), to explore different patterns in daily habit among users, and check if these differences are related with the difference between households at the time scale of the experiment.

Limitations of this study are related to the relatively small number of households, and the lack of more detailed meta-data regarding socio-economic characteristics. The additional information on green contracts, dwelling surface, and residence (Table 2), are available only for part of the households, preventing the possibility of an exhaustive comparison.

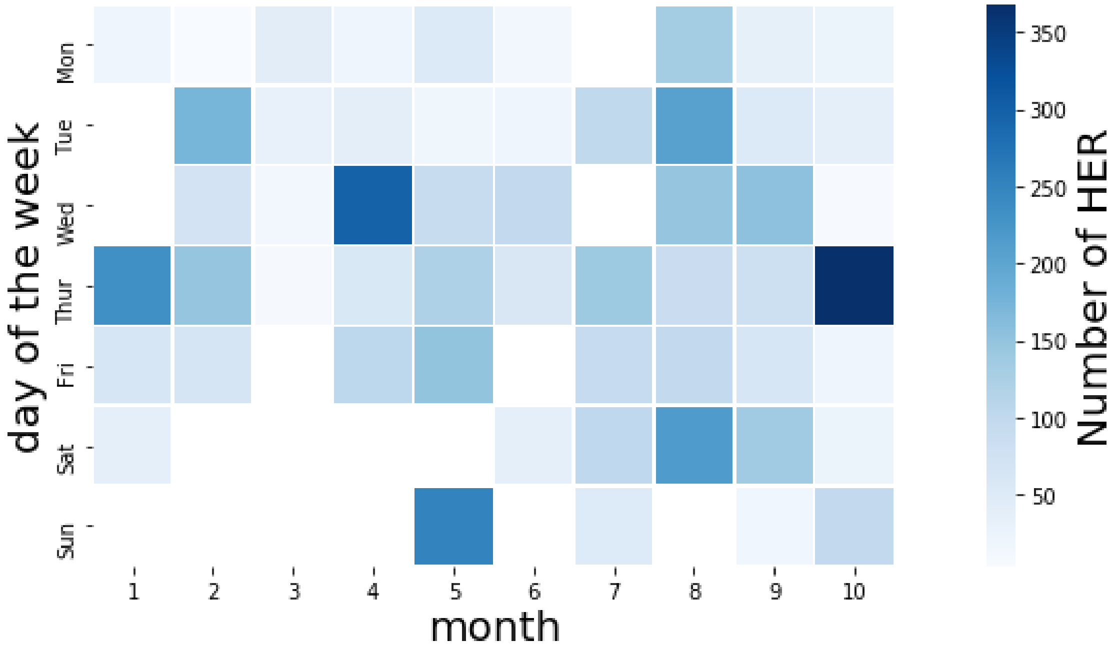
4.2 Energy conservation interventions through Home Energy Report

In the last decades many works studied the effect of behavioral interventions for energy conservation (see Section 2.1), and in particular on the effect of home energy reports (HERs) in reducing residential consumption. Most of the studies though take place in the USA, and residential consumption is often aggregated at monthly level. On one hand, this leads to missing the fine-grained effects of HERs on the time scale of a few days. On the other hand, the variety of behaviors in each country leads to a different potential for energy saving. Here we present a study of the effect of HERs on smart meter data in Italy, in the area of Bologna.

The dataset is the same as in Section 4.1, extended to other towns around Bologna; in fact, the households from this dataset have been periodically (monthly or bi-monthly) receiving home energy reports via email (eHER) throughout all the year. HERs consist of a comparison of the consumption with those of similar clients (similar in terms of type of building and number of occupants), and the comparison with *virtuous* clients. A second part of the report compares the current consumption with the one from the previous year, with information on the $C0_2$ that has been saved/emitted more/less than previous year. Finally, tips for energy saving are provided.

The dataset consists of panel data of power load at hourly resolution for 936 household located in the region around Bologna (Italy), for a time span starting from December 2019 until October 2020. These households have been receiving HER monthly or bimonthly, with different reception time for each family. In Figure 10 we show the distribution of reception times depending on the month and the day of the week. These dates are not uniform in the dataset, and reception dates are not synchronized among users.

Figure 10: HER reception dates depending on the month and day of the week.



Our aim is to evaluate the effect of HER reception on users' consumption. As we mentioned in Section 2.1, usually RCTs are implemented to evaluate efficacy of HER, and this allows to assess the causality of the intervention. In our case though, a control group is not available; we therefore consider different approaches. The first one is a two-way fixed effects regression model (2FE) (Section 4.2.1), which allows to estimate the variability in consumption after the report. However, this approach relies on modeling assumption of linear addictive effects (Imai and Kim 2021) which are not completely justified for panel data of power consumption. We therefore apply also the XGBoost algorithm with the explainable AI tool SHAP to check results accuracy (Section 4.2.2).

4.2.1 Two-Way Fixed Effects Regression Models

The two-way linear fixed effects regression (2FE) is one common approach for estimating causal effects from panel data. In our case, we assume that receiving a HER induces an energy saving within a timescale of a few days after the report, and we aim at evaluating that outcome. We therefore ignore the cumulative effect of receiving many reports while focusing on the short-term higher effect.

We apply a 2FE to estimate the daily load consumption of household as a function of the date difference with HER reception, including four days before and six days after the receiving day. If y_{it} is the KWh consumed by family i, in the time interval t (here we assume the binning is the consumption of a day), the regression is:

$$y_{it} = c + \sum_{k=-4}^{6} \tau_k D_{itk} + d_t + g_i + \epsilon_{it}$$

where $D_{itk} = 1$ if k is the date difference with a report reception day for family i on the day t (and $-4 <= k <= 6$), and $D_{itk} = 0$ otherwise. c is a constant, d_t and g_i are day of the year and household fixed effects, respectively. Standard errors are clustered at the level of household.

In Figure 11 we show the regression coefficientsτ_k. On the second day after the report reception, we see an average decrease in consumption, which is not significant though. The energy saving corresponds to around 1.3% of the average daily consumption in the dataset. Further statistics related to the regression are reported in Table 3.

Figure 11: Regression coefficients τ_k (in KWh) to estimate the effect of HERs on the k days after the report is received. The error bars are standard errors clustered at the level of single household.

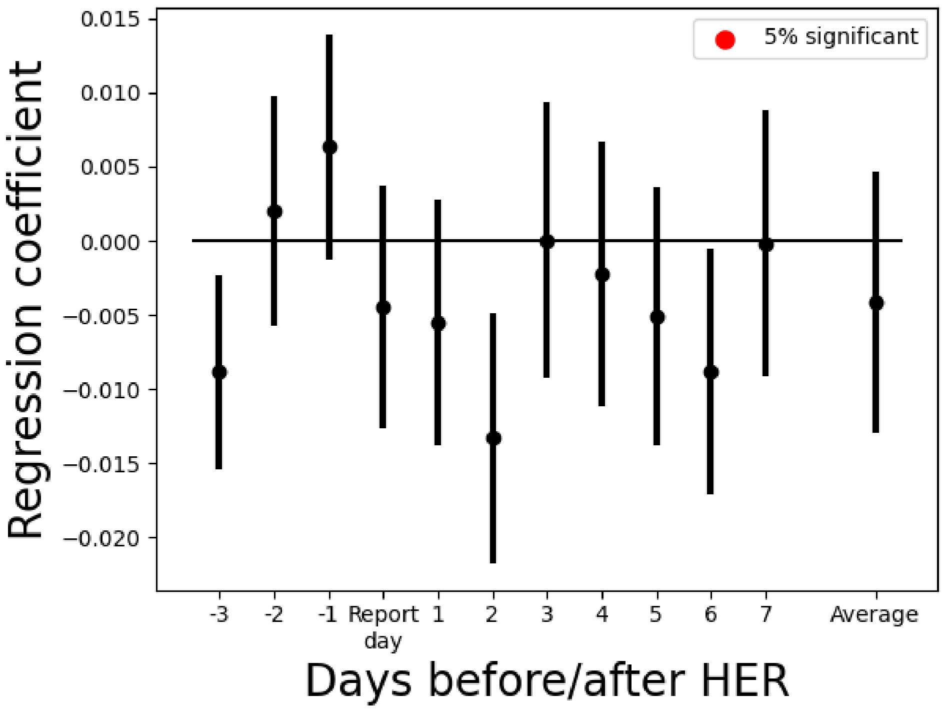


Table 3: 2FE regression summary table.

OLS Regression Results

Model:	OLS	R-squared:	0.638
Method:	Least Squares	Adj. R-squared:	0.637
No. Observations:	301077	F-statistic:	2762
Df Residuals:	299808	Prob (F-statistic):	0.00
Df Model:	1268	Log-Likelihood:	-2.2894e+05
Covariance Type:	cluster	AIC:	4.604e+05
BIC:	4.739e+05		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.9039	0.021	42.654	0.000	0.862	0.945
day -3	-0.0088	0.007	-1.357	0.175	-0.022	0.004
day -2	0.0020	0.008	0.259	0.795	-0.013	0.017
day -1	0.0063	0.008	0.828	0.408	-0.009	0.021
day report	-0.0045	0.008	-0.550	0.582	-0.020	0.011
day 1	-0.0055	0.008	-0.665	0.506	-0.022	0.011
day 2	-0.0133	0.008	-1.581	0.114	-0.030	0.003
day 3	3.649e-05	0.009	0.004	0.997	-0.018	0.018
day 4	-0.0023	0.009	-0.255	0.799	-0.020	0.015
day 5	-0.0051	0.009	-0.591	0.555	-0.022	0.012
day 6	-0.0088	0.008	-1.069	0.285	-0.025	0.007
day 7	-0.0002	0.009	-0.022	0.983	-0.018	0.017
day average	-0.0041	0.009	-0.473	0.636	-0.021	0.013

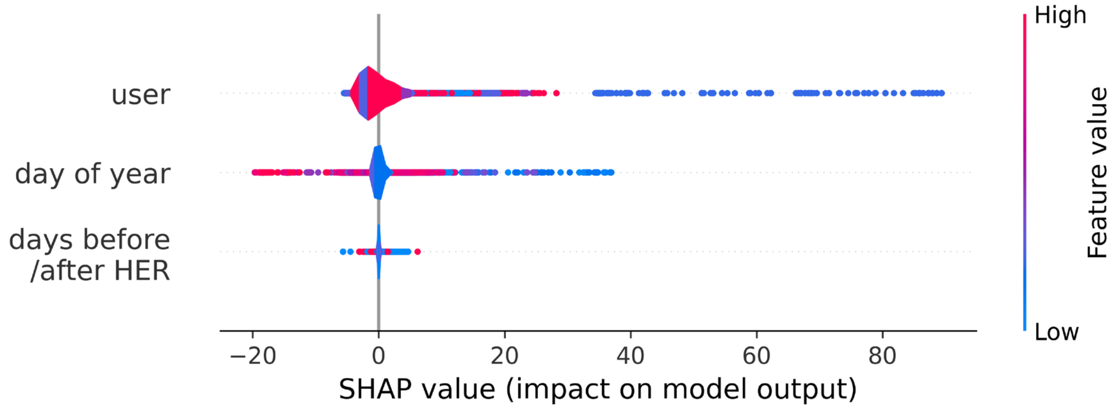
Omnibus:	304985.244	Durbin-Watson:	0.591
Prob(Omnibus):	0.000	Jarque-Bera (JB):	186368313.227
Skew:	4.201	Prob(JB):	0.00
Kurtosis:	124.596	Cond. No.	1.23e+03

4.2.2 XGBoost and SHAP for estimating HER treatment effect

Another approach to estimate the treatment effect of HERs is to build a counterfactual as described in Section 3.1, in order to compare household consumption after HER reception with the estimate in the absence of report. The problem with applying this approach to our data is twofold. On one hand, the staggered reception of reports requires the forecasting to learn from a few weeks period to predict a week of consumption. This is usually not a long enough time when dealing with data with not only weekly, but also yearly periodicity. The second problem come from the specific forecasting of individual power load time series. As we described in Section 3.1.1, power load is a type of time series particularly volatile, and usually reliable forecasting can be made only at an aggregate level, with a historical period of learning data that should span some years.

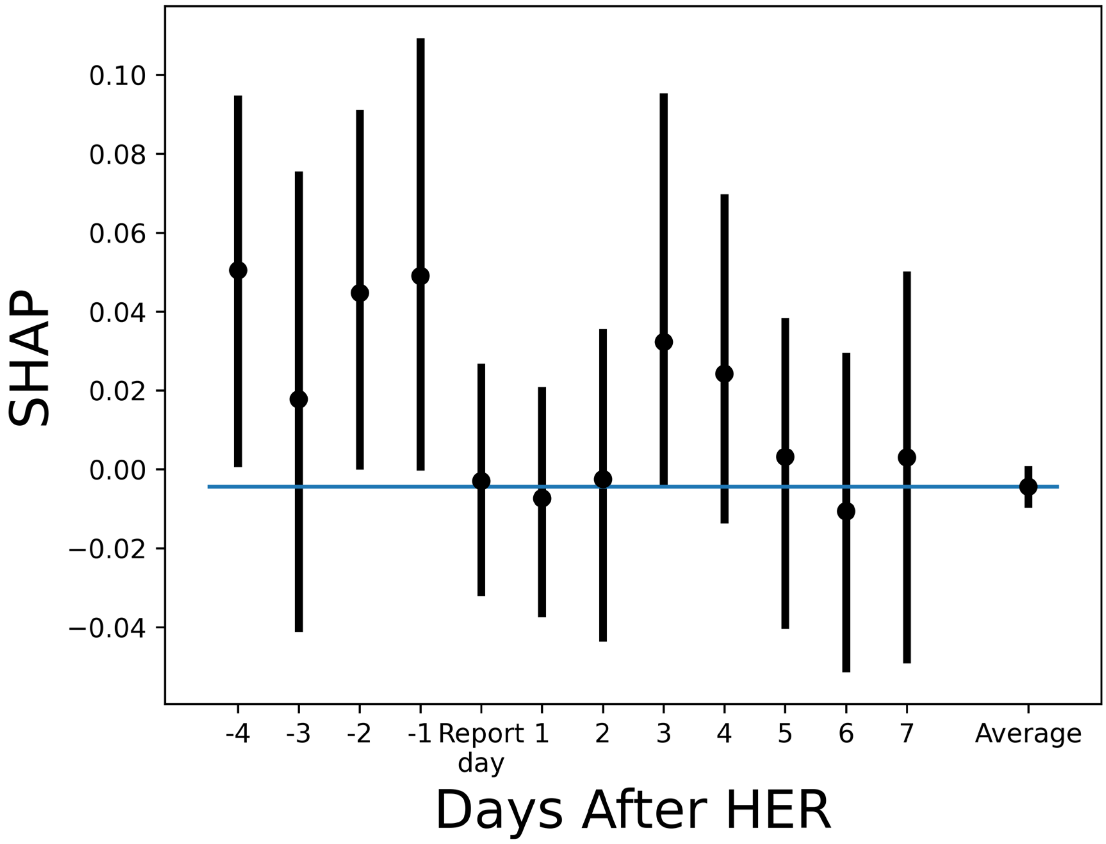
Tree-like algorithms are typically adopted for prediction, and they can be extended to time series prediction. In this case, features can be the month, day of week, day of year, hour etc. When applied to power load data, gradient boosted trees (XGBoost) can be used to build a counterfactual by predicting the values of the consumption, depending on features as day of the week, temperature etc. (Souza 2019). Taking into account the limitation of our dataset in estimating a good enough forecast of time series, we do not use XGBoost to build a counterfactual, whereas we interpret the evaluation of the treatment effect as a feature importance problem. In other words, instead of using XGBoost to predict the counterfactual and then comparing it with the consumption values, we decide to simply fit the whole consumption period and then check the influence of the feature *day before/after HER* on the model (Lundberg et al. 2020). In this way, we build a machine learning extension of the regression model in evaluating the impact of the treatment. To the approach of Section 4.2.1, we assume the effect of HER to be visible only in the week after the reception. We also include as features few days before the HER, with the aim of testing the soundness of our results. Finally, SHAP, an explainable AI (XAI) tool (Rojat et al. 2021) described in Section 3.3, is used to estimate the features importance for the machine learning model.

Figure 12: Feature importance for all the features used in the XGBoost model. Positive (negative) SHAP values indicate an increase (decrease) in the model output (i.e. energy consumption).



The features considered for the model mimic the regression of previous Section 4.2.1. Figure 12 shows the relative importance of the different features used in the XGBoost model to describe the time series. Positive SHAP values indicate that the feature is positively impacting the output of the model, the opposite holds true for negative values. Each point that the model predicts (in our case, each day of consumption of a household) has a different SHAP values for each feature used for prediction. If the feature *day 1 after HER* has a negative SHAP values, it means that *day 1 after HER* is leading to a reduction in consumption for that household. The most important feature is the identity of the household (*user*), followed by the day of year. The day after/before HER influence on the model prediction is shown in Figure 13, where we can see that on the two days after the report reception, there is a decrease in consumption compared to the previous and following days, but from the third day the consumption level is again comparable to the values before the report.

Figure 13: Same dataset as in Figure 11, analyzed with XGBoost and SHAP values. Errors correspond to the 95% quantiles obtained from many bootstrap realizations (50 realizations).



The last week of data (from 18 October 2020) is excluded from the training dataset for model validation. We compute the RMSE (root-mean-square error) between the test data and the model prediction. This value should be compared with the average of the test data in the same time interval, to estimate the relative importance of RMSE. At a daily resolution (Figure 13), we find a RMSE of 3.64 with mean value: 6.95 (RMSE/mean $\approx 52\%$). We also consider a model with data at hourly resolution, therefore including hour of the day ad feature. This model performs worse than the daily resolution one, we get RMSE: 0.217 and mean 0.212 (RMSE/mean $\approx 102\%$). Another approach that we consider is to only model the consumption on the hours between 8am and 19pm. The results are similar to those of the daily consumption.

4.2.3 Conclusions

Our analysis shows a slight decrease in consumption in the days just after HER reception, but the reduction is not significant. Results are obtained by applying both the traditional two-ways panel regression model with fixed effects, as well as the ML algorithm of gradient boosted tree (XGBoost) with SHAP values for features importance estimation. The use of the ML algorithm confirms similar results as the panel regression approach.

Bigger dataset and longer time-series would be helpful in assessing significance of results. Future work will analyze data with these characteristics from the following years. Another difficulty is the absence of a control group, considering that power consumption is subject to common external influence (price increase, lockdown...). Next steps include the study of a similar dataset where a control group is provided. We will then be able to explore the heterogeneity in the response and to provide better support for the statistical significance of the findings.

5. Italian case study 2 - Isernia

This work combines the classical regression approach, as in (Houde et al. 2013), with a machine learning approach, as in (Kwac, Flora, and Rajagopal 2014), to better understand if and how power consumption behavior changed concurrently with the installation of an in-home displays providing real-time feedback in a sample of Italian households. A complementary study on the same dataset involves a functional data analysis to uncover patterns related to appliances ownership (Fontana, Tavoni, and Vantini 2019).

5.1 The in-home display

In 2011, one of the largest electricity companies in Italy started a 3-years-long pilot project in the area of Isernia (mid-south Italy) to test new smart-grid-related technologies and inform future network restructuring plans. One part of the project dealt with customer engagement for demand response. In this context, a kit was distributed to thousands of end users to enable active participation by making people aware of how much electricity they were consuming.

The main interaction with the kit occurs via a display installed in the house, informing users about instantaneous consumption, as well as daily, weekly and monthly summaries. The display also provides information about the current billing slot (days are split in 3 billing periods, peak (F1), intermediate (F2) and off-peak (F3)) and the time at which the next slot will enter into force. If users enter information about their billing tariffs, they also get feedback on monetary expenditures. Users can set goals, and are also warned by an acoustic signal whether their power consumption exceeds the contractual obligation (set at 3kW for most customers).

Features of the in-home display for real-time power consumption information feedback, distributed to households in the province of Isernia.

The display was distributed to residents of Isernia city and surrounding municipalities for free. The company focused on this area for technical reasons, related to the feasibility of high-frequency consumption data measurement and transmission. The display was not randomly allocated: it was first distributed outside the city, and subsequently in the city. The company advertised this opportunity through media campaigns. They faced initial difficulties in recruiting enough volunteers, possibly for concerns about privacy or simply lack of advertisement. They subsequently intensified the promotional campaign, by hosting meetings within local communities, in schools and other public spaces. This had the effect of increasing participation quite rapidly. Overall, the trial was not carried out according to the golden rules of randomized controlled trials. The design is subject to possible self-selection which can hinder the external validity of the program impact. Nonetheless, the trial provides useful information on energy consumption behavior, and is the first one carried out in

Italy on a large sample. We discuss below the measures we have taken to try to mitigate the imperfect (from a research stand point) implementation.

5.2 Methods

The kit, containing the in-home display used for consumption feedback testing, was distributed to thousands of households in the province of Isernia in Italy. Recruitment occurred on a voluntary basis, and was supported through several channels. Between June and September 2012, the kit was promoted with informational days at schools, mass marketing, and collaborations with public authorities and institutions. The official recruitment started in November 2012, while informal tests were running since August. Among those who adhered to the initiative, only a sub-sample, which will be called the "Client" sample, had the in-home display still active at the end of the test period, namely December 2014. This sample excludes non-domestic or non-resident customers, as well as those with a power contract other than 3kW or 4.5kW, as patterns of consumption may be very different in these cases. For each household in the “Client” sample, the utility provided data on monthly energy consumption between January 2012 and December 2014, split by billing time slot[3]. Extra available information includes contractual power and municipality at the moment of joining the program, as well as date of delivery and version of the display. With "Survey" sample, we denote a subset of the "Client" households that agreed to provide also information on the demographics of family members, the number of appliances available in the house, and some characteristics of the dwelling.

For another subset of the "Client" households, which will be referred to as the "Curves" sample, it was possible to obtain higher frequency readings of energy consumed, i.e. every 15 minutes, at least for a fraction of the full 3-year span. The collected load curves are re-sampled to 1-hour time steps for the purposes of this analysis. To filter outliers on the higher end, we assume that electricity can be withdrawn at most with power exceeding 10% the contractual value. Hourly readings exceeding 3.3kWh or 4.95kWh are thus removed for 3kW or 4.5kW contracts households respectively. These thresholds represent the same power levels above which service would generally be discontinued after a while. Still, only a tiny fraction of observations exceeds these extremes, corresponding to around 99.999% quantiles of their respective datasets. A slightly more restrictive threshold is assumed on the lower end, removing data below the 0.1% quantile (i.e., 4Wh). This should filter very low readings which might correspond to either faulty sensors, blackouts or empty houses, i.e., not interesting cases for the analysis. Gaps up to 2 hours of missing data are interpolated linearly from available values. Days remaining with missing points after this

3 Time of use is classified into 3 categories: F1 (on-peak), from Monday to Friday 8am-7pm, national holidays excluded; F2 (intermediate), from Monday to Friday 7am-8am and 7pm-11pm, plus Saturday 7am-11pm; F3 (off-peak), from Monday to Saturday 11pm-7am, plus Sunday and national holidays. Depending on the contract, different time slots may have different prices, with the most popular billing having a higher price for F1 and a lower price both for F2 and F3.

interpolation are removed. There is no clear best sample among the three for our impact evaluation purposes, since those with higher number of observations have also a lower number or different set of variables available. Hence, we decided to include all of the three in the analysis that follows.

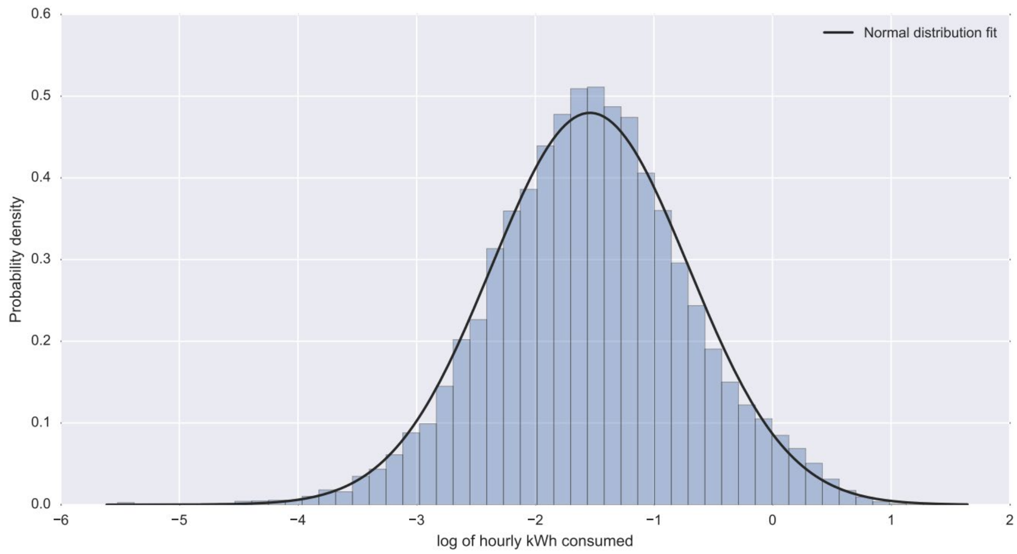
Households who joined and stick to the experiment seem to have consumed much more than the average family in Isernia in 2012, both when considering the estimate from ISTAT of 1973 kWh and the one from TERNA of 2248 kWh. Both average and standard deviation of consumption across households do not change much between the sub-samples considered. When looking at survey data, we observe that the program involved families with above-average number of members and rooms in the house. Ownership of washers, dryers, dishwashers, electric boilers is also above the regional average, while the presence of air conditioning is more marginal. This holds also when considering the intersection of the "Survey" and "Curves" samples. Thus, it appears that the sample of our analysis is not fully representative, pointing to selection bias. The bias was not apparently driven by higher per capita energy consumption: this matches quite well the official statistics of the Isernia municipality. However, families participating in the trial are more numerous, with 3.2 persons per family as opposed to the population average of 2.4. This might be attributed to the initial school campaigns, which were the most successful in getting the program going according to the utility.

Table 4: The different datasets collected during the experiment.

Main Variables	Summary statistics
Client dataset	1834 households
Date of delivery of the display	2012 (2%) 2013 (66%) 2014 (32%)
Version of the display	I (27%) II (73%)
Municipality	Isernia (63%) Other (37%)
Contractual power	3kW (94%) 4.5kW (6%)
Yearly energy consumption, average	2814 kWh in 2012 2739 kWh in 2013 2614 kWh in 2014
Yearly energy consumption, standard deviation	1103 kWh in 2012 1069 kWh in 2013 998 kWh in 2014
Survey dataset	804 households
Number of family members	<3 (26%) 3 (26%) 4 (37%) >4 (10%)
Age class of head of HH	18-30 (3%) 31-50 (37%) 51+ (54%)
Sex of head of HH	Male (85%) Female (14%)
Number of TVs per HH	1 (27%) 2 (31%) 3 (23%) >3 (18%)
Number of fridges/freezers per HH	<2 (32%) 2 (56%) >2 (12%)
Number of ovens/stoves/microwaves per HH	1 (39%) 2 (40%) >2 (16%)
Number of washers/dryers per HH	0 (1%) 1 (84%) >1 (15%)
Number of ACs per HH	0 (81%) 1 (14%) >1 (5%)
Number of boilers per HH	0 (86%) >0 (14%)
Number of heaters per HH	0 (78%) 1 (21%) >1 (2%)
Number of rooms	3 (6%) 28 4 (13%) 5 (26%) 6 (55%)
Load curves dataset	966 households
15min time-step energy consumption	On median, tracked: 62% over 3 years 456 days before display 194 days after display

Hourly energy consumption computed from the "Curves" dataset appears to have a log-normal distribution (see Figure 14), in agreement with the literature (Kwac, Flora, and Rajagopal 2014). It includes 966 households and 26304 time periods, for a total of 14,805,000 non-null observations (i.e. ~60% of all client-hour combinations).

Figure 14: Distribution of hourly power consumption is well approximated by a log-normal distribution.

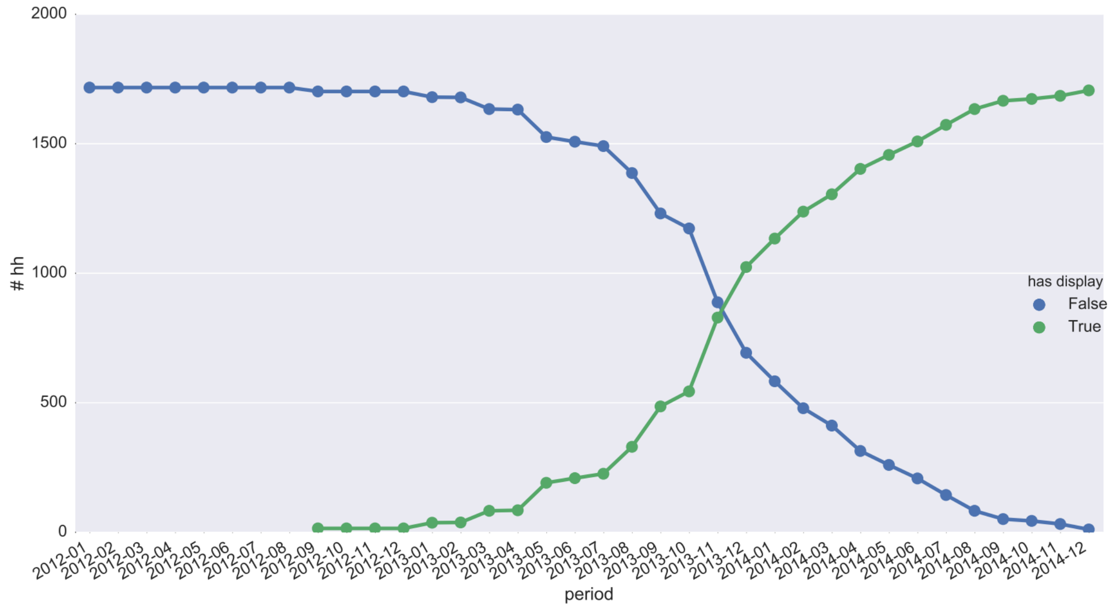


5.3 Regression of daily electricity consumption

First, we quantify the impact of the in-home display on the average power conservation effect. If we were to look only at conditional averages of consumption between those with and without an in-home display over time, we would risk attributing to the display the merit of an already decreasing trend in demand (~3% reduction per year in 2013 and 2014, based on ISTAT and TERNA data). Hence it is important to rely on some other identification strategies.

Impact is usually evaluated against a counterfactual consumption, in this context the hypothetical one of those who already received the display if they hadn't received it. Ideally, a control group is sampled to provide such counterfactual. Since we do not have access to the latter, we exploited as an alternative the gradual phase-in of the experiment, building the counterfactual on the basis of the consumption of those who haven't received the display yet at any point in time. Figure 15 shows how households progressively received the display. Although our identification strategy is not ideal, recent research seems to indicate that high frequency data can be used to estimate causal effects in non-experimental research designs[4]. This identification strategy works as long as the in-home display delivery date is plausibly random with respect to HHs characteristics. Evidence supporting this hypothesis was found by ensuring that the baseline consumption level of a HH, one distinguishing characteristic available for all HHs, was not a predictor of the delivery date itself.

Figure 15: Number of observed clients with and without display at each month in the 3 years of the experiment.



We cast the impact evaluation problem as an ordinary least square (OLS) regression, either in a pooled or fixed effects setting. Observations are defined over the set of households and timestamps, which both depend on the sample considered. Timestamps can be represented either as (month, year) pairs with the low-frequency (LF) monthly datasets, or as (day, month, year) triplets with the high-frequency (HF) sub-hourly dataset. The dependent variable is log of daily power consumption, obtained dividing monthly levels by the number of days in a month, or resampling high-frequency data when available.

Independent variables include:

- presence of the display (either 0=no display yet, or 1=display received);
- day of the week (from 0=Monday to 6=Sunday, only for high-frequency data);
- month (from 0=January to 11=December);

4 See for example, David Rapson work entitled 'Can high-frequency data and non-experimental research designs recover causal effects? Validation using an electricity usage experiment', with Katrina Jessoe and Douglas Miller, presented at 2015 AERE (http://aere.org/summer/documents/AERESummerConference2015Program.pdf)

- year (from 0=2012 to 2=2014);
- municipality of the household;
- household fixed effect;
- time fixed effect (from Jan 2012 to Dec 2014 for LF data, and from 1st Jan 2012 to 31st Dec 2015 for HF data);
- survey variables (family size, average age and sex ownership; appliance ownership; number of rooms in dwelling);
- weather variables (average temperature, both in linear and squared terms).

By combining different choices of dependent and independent variables we come up with 6 plausible models to explain consumption (see Table 5). All models share the same structure:

$$y_{i,t} = \alpha D_{i,t} + \sum_j \beta_j x_{j,i,t} + \varepsilon_{i,t}$$

where D indicates the presence of the display at time t in household i, x_j is one of the other independent variables, possibly dependent only on i, t or both, α and β_j are the corresponding regression coefficients, y is a transformation of daily power consumption according to one of the definitions listed in Table 5, and ε is the error term. The addition of some variables, like the survey ones or the day of the week, imposes restrictions on the usable sample of observations, as explained when introducing the three available datasets. Also, addition of fixed effects prevents the inclusion of covariates related to attributes absorbed by the fixed effects. Intervals of confidence for the coefficients are estimated both with and without a so-called cluster-robust covariance estimator, treating each individual as a cluster. Given the limited sample size, we could not explore potential interactions between covariates. If more data were available, it would be possible to quantify intervention effects per subgroup of households sharing specific attributes.

To better explore the different information sets available for the three different samples, and to test the robustness of results to different timescales, subsamples and model structures, we consider multiple models.

The first two models are fit to the "Client" sample, which is the largest of the three and is balanced over time, although limited to monthly observations. Given the very few observables available other than consumption, we capture the time-independent but HH-dependent unobservable factors driving power demand with HH fixed effects. The difference in the two models is how they capture time-dependent but HH-independent factors: either with time fixed effects (at monthly resolution, from Jan 2012 to Dec 2014) or with a 2-degree polynomial of average temperature plus monthly and yearly dummies. Historical temperature statistics are computed from local weather stations records, which cover only part of the full-time span of the experiment and hence slightly limit the number of observations.

The second two models focus on the “Load” sample, which covers fewer households than the “Client” sample but it introduces high frequency measurements. The main difference with the first two models is that time-dependent variables have a daily resolution. The daily fixed effect model spans a period from 1st Jan 2012 to 31st Dec 2014. When not including time-fixed effects, we add to the time-dependent variables above also day-of-the-week dummies.

The fifth model is fitted to the “Survey” sample, which is also smaller than the “Client” sample but allows for correlating power consumption with HHs surveyed characteristics. We replace the HH-fixed effects with these variables, which include family size, socio-demographics, appliance ownership and number of rooms in dwelling.

In the last model we change the dependent variable: the regression of daily consumption is set up as in model (2), but it excludes nights and weekends (i.e. it focuses on the F1 billed time slot). This is when most of the energy-intensive household activity tends to take place.

The choice of these models reflects the varied nature of the dataset. Model (1), (2) and (6) cover the highest number of HHs, but they have the least amount of covariates and coarsest time resolution. Model (3), (4) and (5) cover fewer HHs, but provide further details on sub-monthly consumption and on the peculiarity of each HH. Model (2) and (4) offer plausible alternatives on the role of time in affecting consumption, modeled either via weather effects combined with seasonality dummies, or via time-fixed effects. The former is more physically grounded, while the latter is not subject to the limits in quality and availability of weather data. Lastly, model (6) taps into possible on-peak effects, versus the overall effect on which the other models focus. Given all these trade-offs, a single-model assessment risks missing the nuances of our dataset and to provide a less robust assessment.

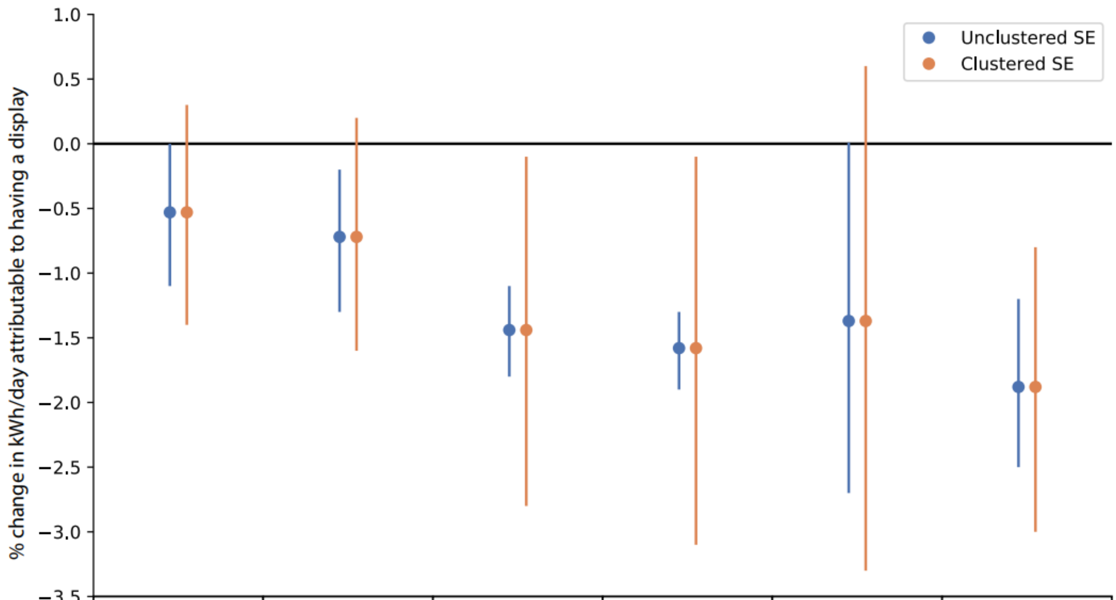
Table 5: Models used for estimating the impact of having a display on daily power consumption.

Model name	(1) CLIENT	(2) CLIENT Month FE	(3) LOAD	(4) LOAD Day FE	(5) SURVEY	(6) CLIENT F1
Sample						
Client	x	x				x
Survey					x	
Curves			x	x		
Dependent variable						
log(daily consumption from monthly data)	x	x			x	
log(daily on-peak consumption from monthly data)						x
log(daily consumption from high-freq data)			x	x		
Independent variables						
Display received	x	x	x	x	x	x
Municipality					x	
Year dummies	x		x		x	
Month dummies	x		x		x	
Weekday dummies			x			
Avg. Temp. + (Avg. Temp.)2	x		x			
HH fixed effects	x	x	x	x		x
Month/Year fixed effects		x				x
Day/Month/Year fixed effects				x		
Survey variables					x	

For each model we run a separate OLS regression. These regressions imply a comparison at each time period between the sets of households with and without an in-home display.

The regression coefficient of primary interest is the one related to having received a display (i.e. α). This coefficient represents the average percentage increase in consumption due to the presence of an in-home display (Figure 16).

Figure 16: Percentage change (positive = increase, negative = decrease) in daily power consumption attributed to having received the in-home display.



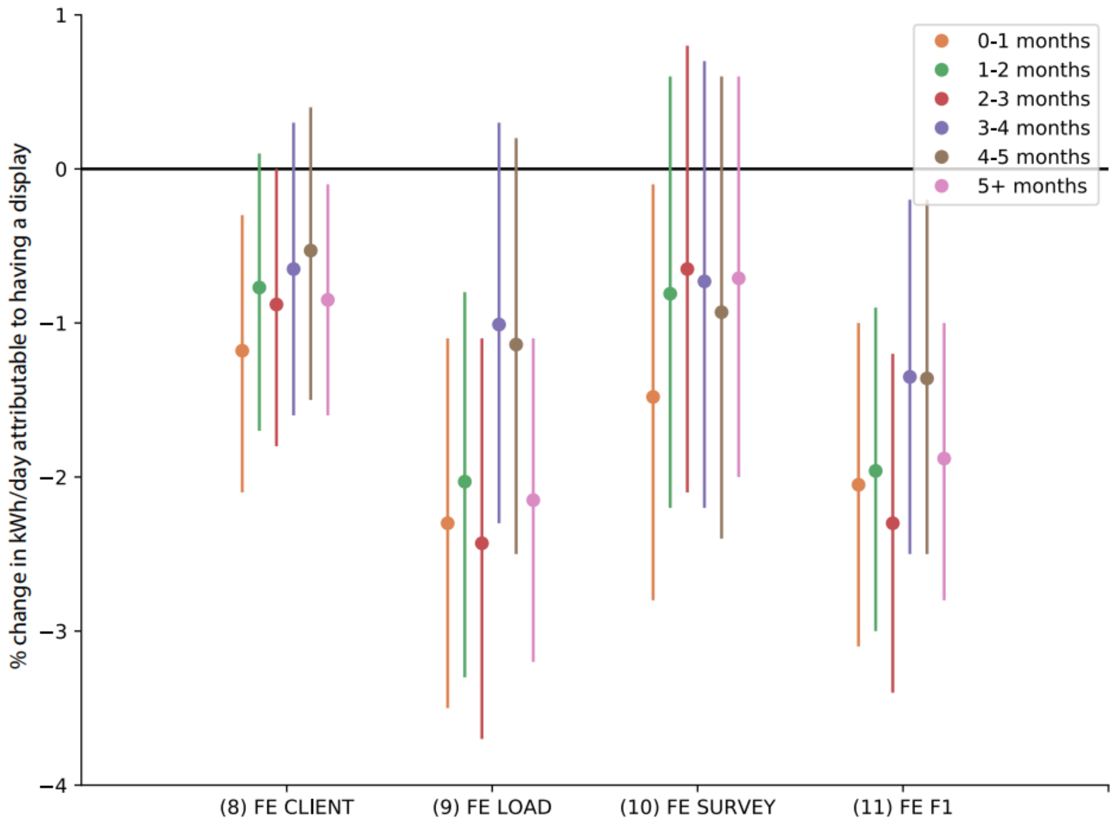
	(1) CLIENT	(2) CLIENT Month FE	(3) LOAD	(4) LOAD Day FE	(5) SURVEY	(6) CLIENT F1
Effect	-0.5%	-0.7%	-1.4%	-1.6%	-1.4%	-1.9%
N. Obs.	42e3	59e3	372e3	513e3	24e3	59e3
Adj. R^2	0.79	0.77	0.61	0.60	0.34	0.72

The estimated effect has the expected direction. Although small, it is statistically significant for half of the models, and the magnitude is in line with the most recent studies. It is not significant for the models involving the client and the survey datasets, while the effect seems greater for the load curves dataset. This latter sample may represent households with higher consumption per capita in the first place, hence with more options to reduce consumption. In the models based on HF data, unclustered uncertainty ranges shrunk considerably thanks to the abundance of observations, while the clustered ranges take into account the correlations between observations and the masking of uncertainty is less evident.

Such a small average conservation effect can have several explanations. The in-home display alone did not provide any monetary incentive or direct message promoting energy conservation, which are documented to be more effective means than information feedback alone. The geographical area of interest is relatively mountainous, with a climate not requiring air conditioning. Heating on the other side is rarely done via electricity. Power demand is mostly related to lighting and other low-consumption appliances, making it harder to save more energy. Nonetheless, even a 1% reduction, if brought to a large scale may matter to the power industry.

A question which recurs in this literature is how long the effect of feedback lasts over time. We focus on the models with fixed effects for households and time periods, as well as variables controlling for having received the display within the last month (0-1 months), more than 1 month ago but less than 2 (1-2 months), and so on (Figure 17). The first three models are fitted to the three different samples mentioned. The last one excludes consumption in nights and weekends. Looking at the resulting regression coefficients, we do not find a clear decreasing trend in effectiveness of the display as months pass by. For three models out of four, the expected reduction in consumption reduces either after one or three months of having the display. For the client model the effect is overall smaller and the reduction over time is less apparent. Overall, an average declining trend in reduction is observable in most cases, but it is not statistically significant for a p-value threshold of 5%.

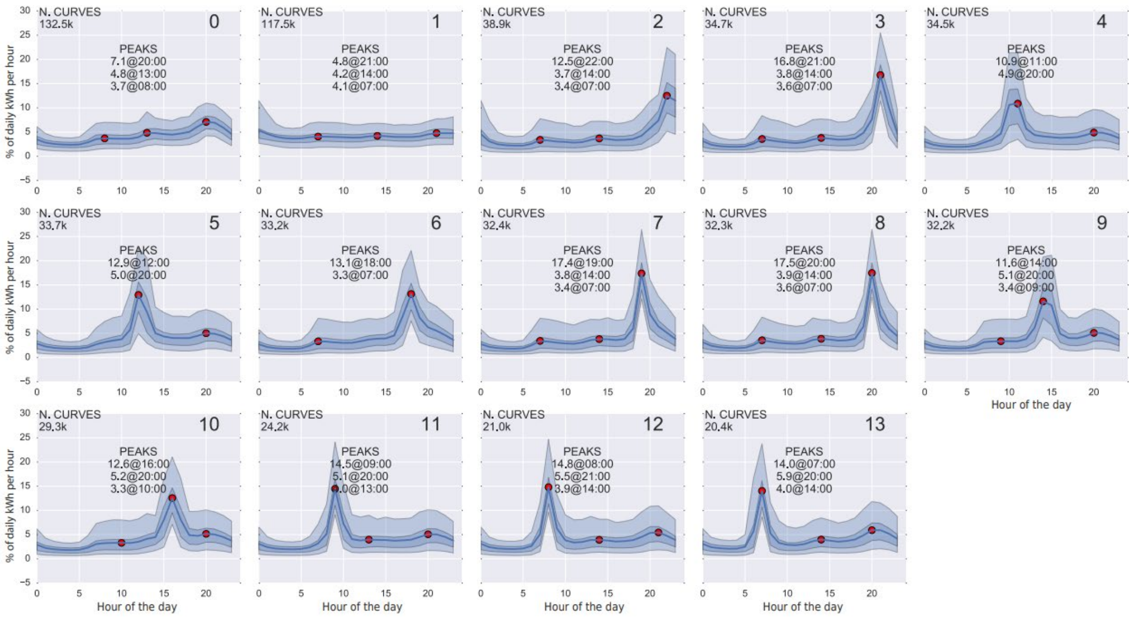
Figure 17: Effect of the in-home display over time.



5.4 Clustering of load shapes

We discuss the clustering for this case study in greater detail in newTRENDs deliverable 5.1, and we refer the reader to that report for further details. For convenience, we report the 14 prototypical daily patterns of power consumption in Figure 18, characterized by different number, timing and intensity of peaks throughout the day.

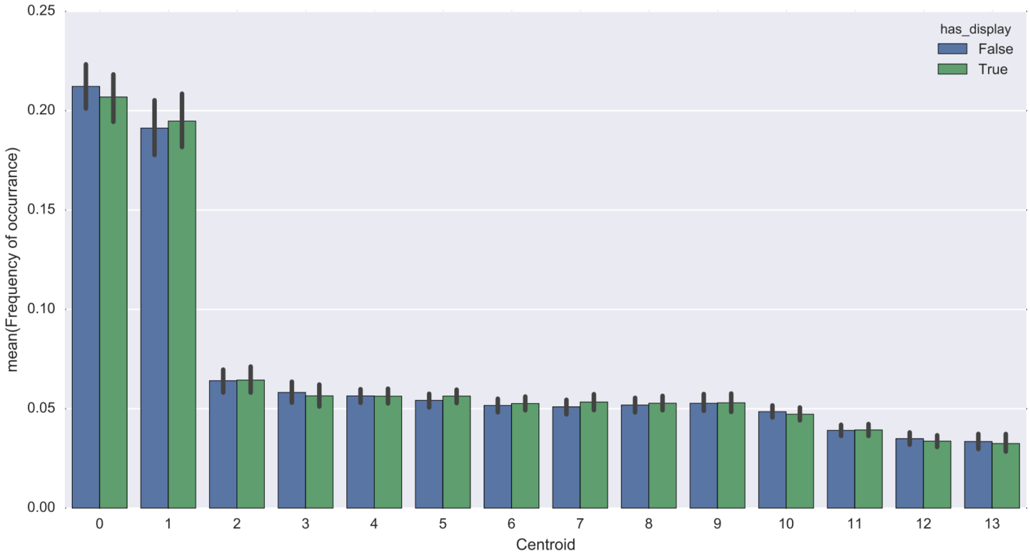
Figure 18: Prototypical load shapes resulting from the K-Means clustering. Primary and secondary peaks are described in each subplot, in terms of % level of daily power consumption and hour of occurrence.



After having identified the centroids, frequencies of occurrence of centroids are calculated for each household, distinguishing the days before and after the arrival of the in-home display. 70 days of data are required to build such frequency vectors. If any behavioral change happened in the patterns of daily consumption due to the in-home display, this should be reflected in the before-after difference of such vectors.

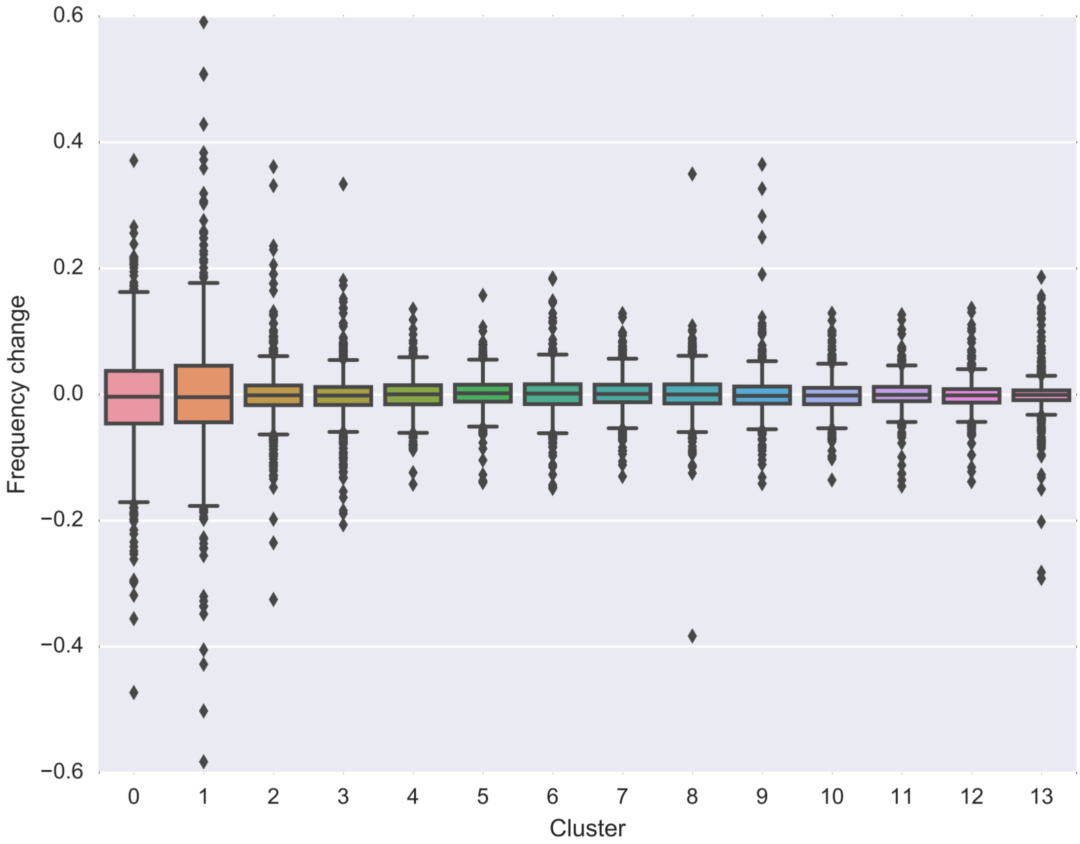
For each household, frequencies of occurrence of centroids are calculated for the days before and after the arrival of the in-home display. If there was a common trend in behavioral change due to the in-home display, we should see a significant change in such frequencies, moving away from some representative shape towards other ones. As shown in Figure 19, this is not emerging from the data. The average behavior of the sampled clients, as coded in these vectors of frequencies, does not change significantly conditionally on having or not a display.

Figure 19: Frequencies of occurrence of centroids, averaged across clients, to which load shapes before and after the in-home display arrival are clustered. Black vertical bars represent confidence interval for the means.



Nonetheless, if we plot the change in frequencies before and after the delivery of the in-home display for each client, a wide heterogeneity emerges (Figure 20). Most households seem to have maintained stable consumption patterns over time, as indicated by the mass of the distributions of frequency gravitating around 0. Still, several households exhibit much more flexibility in consumption, as shown by the long tails of some of the 14 distributions. Overall, the in-home display was not consistently able to provide the incentives required for an average visible shift of peaks over time, but it did work for a subset of the population. This result is consistent with the literature emphasizing the high heterogeneity of the impact of real-time feedback programs.

Figure 20: Changes in frequencies of occurrence of centroids, to which load shapes before and after the in-home display arrival are clustered. Each household corresponds to a series of 14 points, one per cluster.



5.5 Conclusions

According to our estimates, which involve 6 alternative plausible statistical models, an average reduction in daily power consumption of 0.5-1.9% can be attributed to such device, even though with marginal statistical significance. This is in line with more recent experimental studies on the topic.

Average hourly power consumption behavior, measured in terms of distributions of preferences for representative load shapes by each user, seems to remain unchanged before and after the arrival of the display. Nonetheless, a wide heterogeneity has yet to be explained.

Further data on the households would be useful to try to segment the analysis in meaningful groups, identifying those customers for which information feedback led to either significantly lower consumption levels or significantly different consumption habits.

Consumption data at the appliance level would provide invaluable information on behavior and behavioral change, even though at the cost of increasing computational tractability.

Results are specific to the group involved in the study, and are not easily generalizable to the wider population. The trial was carried out in an area of the country with lower than average per capita energy consumption, as well as income, allowing for more limited adjustments and investments. On the other hand, the sample households appear to have more family members than average, possibly young children. Future better experimental designs will also help to avoid issues of sample representativeness and absence of control group.

All these potential steps for further research could lead to a better understanding of how such interventions perform at large scale, and whom these interventions could be most beneficial for.

6. Polish case study

6.1 Introduction

Different households have different propensity to save energy (Long, Mills, and Schleich 2018). This observation leads to a conclusion that for some households, behavioral measures will not have a positive influence on their energy consumption, while for others the observed impact could be significant. A rational policy choice would be to allocate public resources (e.g., funding for a behavioral change program) to households, who are likely to benefit from the intervention. An important research question arises: how to consider the different effectiveness rates of behavioral change measures for different groups of households?

In the Polish case study we investigate the impacts of a behavior change program on energy savings in clusters of households from treatment and control groups (i.e., households who participated in the program and those who did not participate in the program), using unsupervised machine learning based on k-means algorithm. As a result, we provide smart-meter data-based inputs for energy-demand modeling, i.e., profiles of hourly differences between the energy consumption in different clusters of households – both participating and not participating in the behavior change program aimed at energy savings. Our results inform energy models of common barriers to modeling of energy consumptions behaviors, such as variations in responses of households to programs aimed at changing residential energy consumption patterns. We also propose potential ways for overcoming these barriers, through modeling of behavior change measures taking into consideration their different impacts on different groups of households.

6.2 Methods and data

The dataset from Poland includes hourly smart meters readings of electricity consumption from 1489 households, from March 1st-31st, 2017, i.e. during the behavioral change program implemented between September 2016 and November 2017. 1,271 households participated in the experiment and 218 were the control group. The socio-economic characteristics of both groups are similar in terms of the age, apartment size, and number of individuals living in a household. Households were recruited during the “door-to-door” campaign, i.e. they were visited by interviewers and asked for a consent to participate in the experiment and share their electricity consumption data. The dataset covers only apartments located in blocks of flats (no single-family houses). In all flats, space and domestic hot water are heated from the district heating network, which translates to the fact that electricity is not used for those purposes. Also, air conditioning is rather rare. For cooking, mainly natural gas is used, however, it cannot be excluded that some of these flats are equipped with electric cookers and ovens. Electricity is mainly used for lighting and powering of appliances.

Within the behavioral change program “Step by step commitments to energy savings”, the targeted households were regularly contacted through email or by phone and encouraged to try new ecological changes in their behavior related to energy consumption. The encouragement went beyond standard one-way communication techniques, as households were asked to make a commitment to perform specific self-selected tasks and after two weeks were asked if they succeeded and depending on the answer, another or the same activity was proposed to strengthen the initial commitment.

6.3 Results

All statistical features (minimum value, 1st quartile, median, mean, 3rd quartile, and maximum value) for the daily consumption (00:00-23:59) have lower values in the treatment group (Table 6), compared to the control group (Table 7), however differences between values observed at night (23:00-4:59) are not statistically significant (p-value > 0.05).

The impact of the program on energy consumption resulting from the behavioral intervention seems to differ throughout the day (Figure 21 and Figure 22). Negligible changes (statistically insignificant, i.e., p-vale > 0.05) are observed at night (23:00-4:59). This suggests that the behavioral change program had no or minimal effect on the baseline energy consumption related to using appliances that require continuous powering such as a fridge. Starting from 5:00, energy savings starts to grow. In fact, the biggest savings, are observed in the morning (between 5:00 and 7:59) and in the evening (16:00-21:59) and are significantly lower in the middle of the day (8:00-15:59) and at late evening (22:00-22:59). This reflects the regular activity level of a typical household, which spends most of the day outside (work, school etc.), leaving its apartment in the morning and coming back in the afternoon. On average, energy savings in the Treatment Group reached 4.19% (±2.71pp), compared to the Control Group.

Despite the total daily consumption decreases, there are periods when hourly consumption grows, especially the 1st quartile (e.g., at 3:00, 10:00, 22:00, 23:00). For most hours the highest decrease is observed for 3rd quartile, which suggests that the behavioral program had greater effect of households consuming more energy, while it is hard to reduce energy consumption in household where it is already low only with soft measures such as the investigated behavioral program.

While mean and quartile values show tendencies of a group participating in the experiment, minimum and maximum values reflect behavior of individual households. There is almost no statistically significant difference between the treatment and the control group in terms of minimum energy consumption (Figure 23). In both groups, for all hours, the minimum consumption does not exceed 0.008 kWh, which is very low. On the other hand, the maximum energy consumption in the treatment group is higher in all hours but five (8:00, 10:00, 15:00, 20:00, and 23:00) (Figure 23). This might suggest that there are households that are resistant to behavioral change program and despite the proposed measures, their energy consumption not only does not drop, but even grows. Another possible explanation is that even if a decrease of energy consumption is observed in a longer period such as a full day, local increases of energy consumption in selected hours occur.

Table 6: Statistics of the daily electricity consumption – Treatment Group [kWh].

Hour	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
00:00	0.001	0.068	0.107	0.143	0.174	2.990
01:00	0.001	0.060	0.091	0.116	0.138	1.826
02:00	0.001	0.058	0.086	0.102	0.124	1.285
03:00	0.001	0.057	0.083	0.097	0.120	1.493
04:00	0.001	0.057	0.085	0.099	0.123	1.502
05:00	0.001	0.061	0.093	0.116	0.138	1.716
06:00	0.002	0.069	0.110	0.147	0.177	1.771
07:00	0.001	0.079	0.126	0.172	0.204	2.583
08:00	0.001	0.084	0.136	0.189	0.221	2.437
09:00	0.001	0.085	0.141	0.200	0.235	2.911
10:00	0.002	0.083	0.140	0.200	0.235	2.500
11:00	0.003	0.080	0.135	0.196	0.228	2.624
12:00	0.002	0.079	0.135	0.196	0.227	2.692
13:00	0.003	0.081	0.138	0.201	0.234	3.052
14:00	0.003	0.086	0.145	0.209	0.243	3.717
15:00	0.003	0.092	0.155	0.219	0.259	2.633
16:00	0.001	0.104	0.176	0.239	0.286	3.788
17:00	0.003	0.127	0.211	0.276	0.337	2.903
18:00	0.003	0.159	0.256	0.320	0.399	3.062
19:00	0.007	0.179	0.278	0.340	0.425	2.936
20:00	0.004	0.182	0.275	0.334	0.415	2.710
21:00	0.003	0.163	0.249	0.301	0.372	2.844
22:00	0.002	0.126	0.203	0.249	0.311	2.846
23:00	0.001	0.087	0.145	0.189	0.239	2.349
00:00 – 23:59	1.114	3.171	4.391	4.851	6.043	14.592

Table 7: Statistics of the daily electricity consumption – Control Group [kWh].

Hour	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
00:00	0.004	0.069	0.107	0.144	0.175	1.752
01:00	0.004	0.062	0.092	0.114	0.138	1.547
02:00	0.004	0.058	0.086	0.102	0.126	1.119
03:00	0.003	0.056	0.083	0.097	0.119	1.077
04:00	0.003	0.057	0.085	0.100	0.124	0.926

Hour	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
05:00	0.004	0.063	0.095	0.122	0.145	1.685
06:00	0.001	0.074	0.116	0.159	0.195	1.688
07:00	0.004	0.083	0.134	0.181	0.215	1.745
08:00	0.003	0.084	0.140	0.195	0.234	2.805
09:00	0.002	0.085	0.145	0.210	0.247	2.629
10:00	0.001	0.082	0.142	0.205	0.246	2.606
11:00	0.002	0.081	0.137	0.201	0.242	1.901
12:00	0.002	0.081	0.137	0.200	0.239	2.379
13:00	0.001	0.082	0.141	0.208	0.242	2.489
14:00	0.003	0.087	0.147	0.215	0.253	3.324
15:00	0.004	0.092	0.158	0.226	0.271	2.793
16:00	0.002	0.105	0.183	0.253	0.304	2.653
17:00	0.006	0.131	0.223	0.295	0.361	2.794
18:00	0.008	0.168	0.270	0.340	0.424	2.547
19:00	0.008	0.187	0.291	0.355	0.441	2.576
20:00	0.007	0.187	0.285	0.346	0.427	3.413
21:00	0.006	0.166	0.257	0.313	0.390	2.582
22:00	0.004	0.125	0.206	0.255	0.326	2.335
23:00	0.001	0.086	0.142	0.189	0.243	2.395
00:00 – 23:59	1.115	3.339	4.631	5.023	6.283	14.599

Figure 1: Difference of statistical features between treatment and control group [kWh].

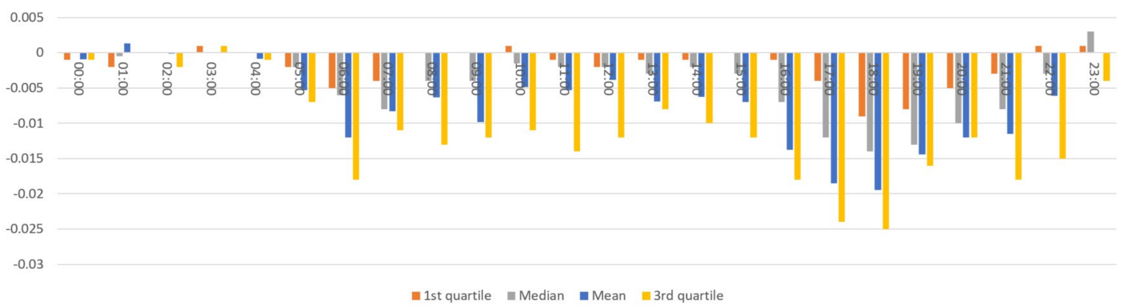


Figure 2: Difference of statistical features between treatment and control group [%].

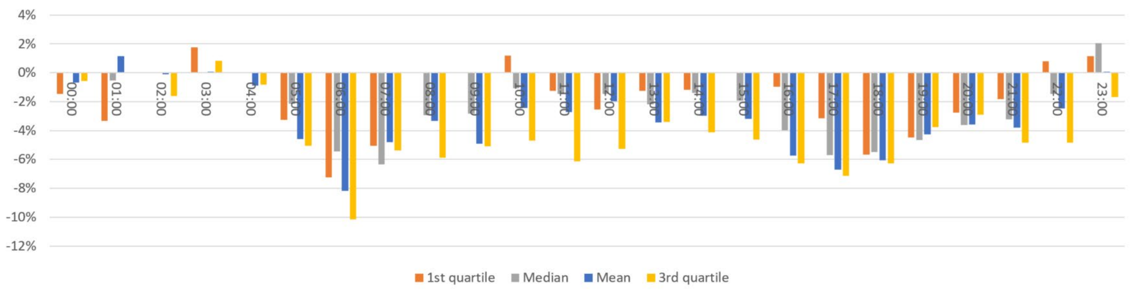
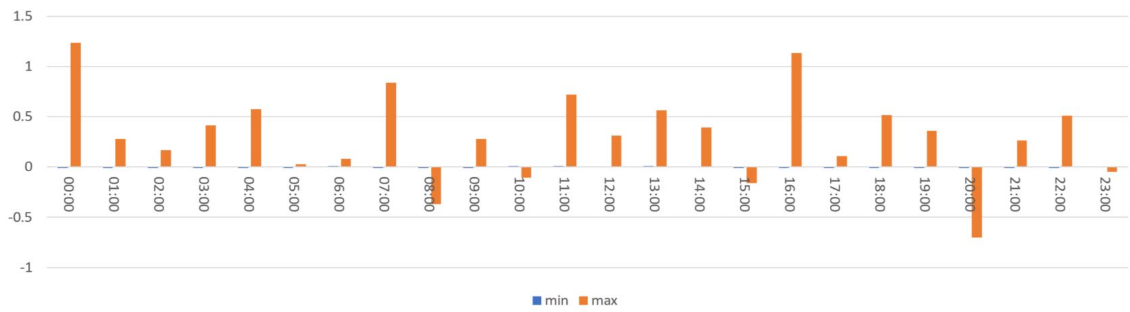
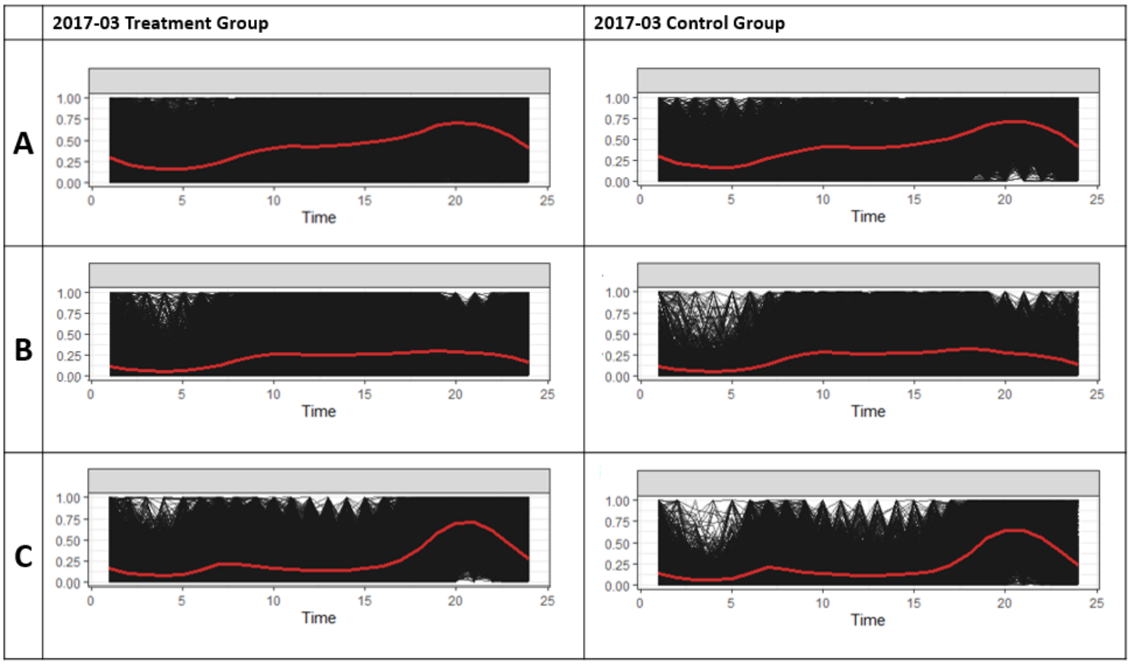


Figure 3: Difference of statistical features between treatment and control group [kWh].



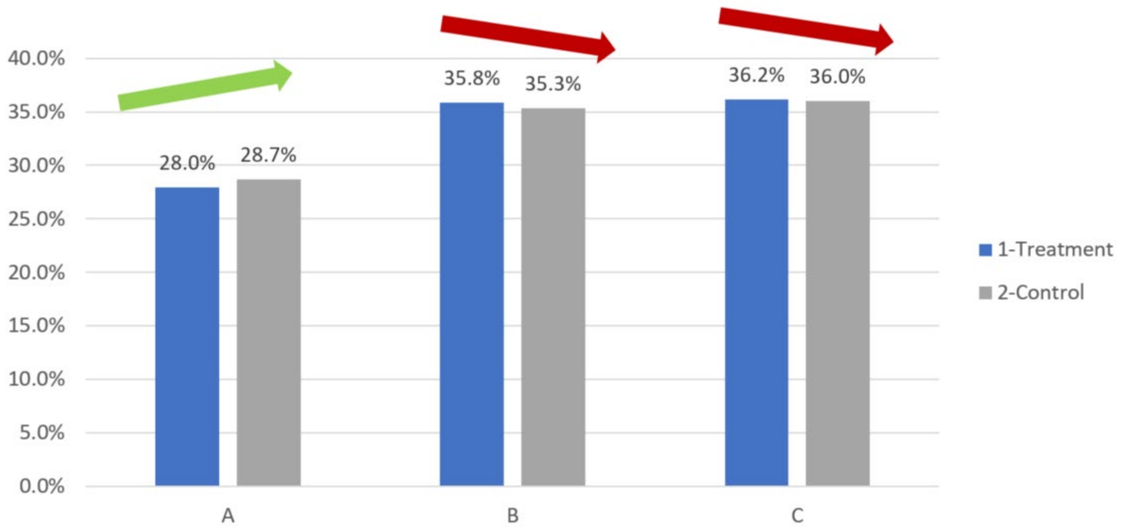
Clustering of the daily electricity consumption, following to the approach presented in the previous deliverable of newTRENDs project, Deliverable D5.1 (Marangoni et al. 2022), shows three distinct consumption patterns (Figure 24). Cluster A is characterized by low consumption at night and high in day, with morning (10:00) and evening (20:00) peaks. Cluster B is characterized by low consumption throughout the day, and even lower in night. No distinct peaks are observed. Cluster C is characterized by moderate consumption in the night and in the morning, with the energy consumption starting to rise at 16:00 and reaching its peak at 21:00. It is important to mention that assignment to the cluster is done based on a daily consumption pattern, so it can change on daily basis. In fact, each household from both treatment and control groups was assigned to each cluster at least once.

Figure 4: Clusters obtained with k-means for the Polish dataset in March 2017. The y-axis represents normalized energy consumption, and x-axis – subsequent hours of a day.



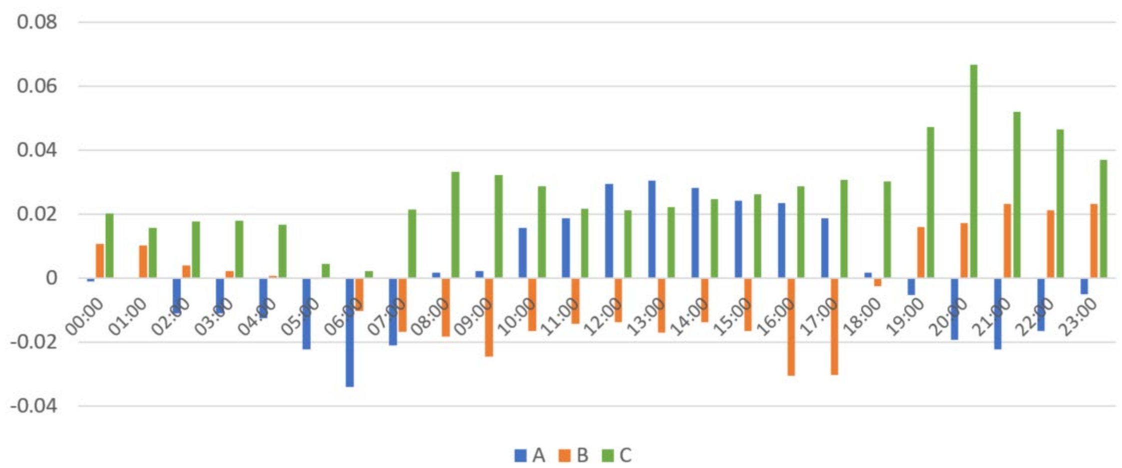
In the treatment group, Cluster C occurs most often (36.2%), and is followed by Cluster B with a slightly lower share (35.8%) (Figure 25). Cluster A is the least frequent (28.0%). In the control group shares of clusters C and B decreases in favor of Cluster A (36.0%, 35.3%, 28.7%, respectively), however, the order remains. The observed shift from Cluster C to A suggests that due to the intervention electricity is consumed in more steady way – the evening peak is shaved, while the afternoon consumption rises.

Figure 5: Cluster frequency in treatment group, before and after behavioral intervention.



At a first glance, there are no significant differences between the load profile of treatment and control groups, neither in terms of the peak amplitude nor time of their occurrence (Figure 24). Hourly analysis, however, shows that there are variations between two groups within clusters (Figure 26). In cluster A we can observe a peak shaving in the treatment group, compared to the control group – energy consumption drops at night and in early morning (up to 7:59), then rises until 18:59, and again shrinks until midnight. The opposite situation is observed in cluster B. Energy consumption lowers between 6:00 and 18:59 and grows in other periods. Interestingly, in cluster C the energy consumption unequally rises the whole day, the most dynamically in the evening (between 18:00 and 20:59).

Figure 6: Difference in the clusters shape between treatment and control group [unit: none (normalized data)].



6.4 Conclusions

In this case study we provide a smart meter-based quantitative assessment of differences between the energy consumption in several clusters of households – both participating and not participating in the behavior change program aimed at energy savings. These results, especially the presented differences in daily energy consumption profiles, inform energy models of the potential variations in responses of different households to programs aimed at changing residential energy consumption patterns. Consideration of these differences should result in a better reflection of household behavior in the modeling of behavioral change programs' impact on energy demand in the residential sector.

7. Conclusions

In this report we explore the effect of behavioral interventions on households' power consumption using a variety of approaches including novel ML techniques. In particular, we extensively use unsupervised clustering algorithms (e.g. K-means) to assess the impact of Home Energy Reports, COVID lockdowns and in-home displays for real-time consumption feedback in terms of variations of daily patterns of power consumption. We identify the effectiveness of the treatment, using both traditional regression and clustering. We show how clustering allows us to get more insights into the impact of the interventions, in terms of change in daily patterns of consumption. The detected reduction in consumption from behavioral intervention was in general low, and more household data are necessary to properly assess the response within specific household subgroups.

In the section on methodology (section 3), we describe multiple ways in which ML could be useful for assessing the impact of interventions on power consumption via smart meters data, and in a few cases these novel approaches have already been successfully applied. ML could be helpful to improve impact evaluation and to overcome limitations, but in general appropriate data availability is crucial. For example, in cases in which a control group is not available, ML techniques are very valuable to build counterfactuals by forecasting time series (Prest et al. 2021). A necessary condition to forecast time series is the availability of sufficiently long pre-treatment period, covering at least one year for analyses involving yearly trends. Another important application of ML is the evaluation of heterogeneous treatment effects, where methods such as causal forests give accurate and unbiased assessment of the CATE (Andor et al.2020). Also in this case, it is important to have enough households' data for each combination of household characteristics.

Starting from the work presented in this report, we can envision different steps ahead, especially within the research agenda of the newTRENDs project. First, as novel smart-meter data become available from utilities, the methods presented could be applied to detect energy demand responses to interventions implemented by the utilities themselves, or to relevant events that might have shocked energy consumption behavior in the recent past, like the stark increase in electricity prices, as well as policies and trends related to the pandemic or the energy sector. These analyses might inform future energy scenarios, where for example prices continue to raise, remote work becomes wide-spread, or smart grids, digitalization, prosumaging, and electric vehicles affect more and more how households consume power. Second, case studies from different regions and projects (e.g. the WHY project) could be brought together for a joint analysis, reflecting on how similarly designed behavioral interventions might yield different results in different regions, and what barriers and success factors drive the outcome of different interventions. Lastly, inputs from this task should also feedback into the modeling stream of work, helping modelers with validation and calibration of household consumption behaviors, and providing sensible ranges of expected energy demand changes attributable to behavioral levers.

The increasing scientific research on these topics and the development of machine learning methods addressing impact evaluation are helping to get the greatest possible value out of the large data ensembles available today. This in turn supports modelers and policy makers to successfully navigate the many possible assumptions in deriving empirically informed policies from those ensembles.

8. References

Abdeen, Ahmed, Farzam Kharvari, William O'Brien, and Burak Gunay. 2021. "The Impact of the COVID-19 on Households' Hourly Electricity Consumption in Canada." *Energy and Buildings* 250: 111280.

Abrahamse, Wokje, and Linda Steg. 2009. "How Do Socio-Demographic and Psychological Factors Relate to Households' Direct and Indirect Energy Use and Savings?" *Journal of Economic Psychology* 30 (5): 711–20. https://doi.org/10.1016/j.joep.2009.05.006.

Abrell, Jan, Mirjam Kosch, and Sebastian Rausch. 2021. "How Effective Is Carbon Pricing? A Machine Learning Approach to Policy Evaluation." ZEW Discussion Papers.

Akhlaghi, Yousef Golizadeh, Koorosh Aslansefat, Xudong Zhao, Saba Sadati, Ali Badiei, Xin Xiao, Samson Shittu, Yi Fan, and Xiaoli Ma. 2021. "Hourly Performance Forecast of a Dew Point Cooler Using Explainable Artificial Intelligence and Evolutionary Optimisations by 2050." *Applied Energy* 281: 116062.

Allcott, H., and S. Mullainathan. 2010. "Behavior and Energy Policy." *Science* 327 (5970): 1204–5. https://doi.org/10.1126/science.1180775.

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9-10): 1082–95.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3): 1117–65.

Allcott, Hunt, and Todd Rogers. 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10): 3003–37.

Almazrouee, Abdulla I, Abdullah M Almeshal, Abdulrahman S Almutairi, Mohammad R Alenezi, and Saleh N Alhajeri. 2020. "Long-Term Forecasting of Electrical Loads in Kuwait Using Prophet and Holt–Winters Models." *Applied Sciences* 10 (16): 5627.

Almazrouee, Abdulla I, Abdullah M Almeshal, Abdulrahman S Almutairi, Mohammad R Alenezi, Saleh N Alhajeri, and Faisal M Alshammari. 2020. "Forecasting of Electrical Generation Using Prophet and Multiple Seasonality of Holt–Winters Models: A Case Study of Kuwait." Applied Sciences 10 (23): 8412.

Alzubi, Jafar, Anand Nayyar, and Akshi Kumar. 2018. "Machine learning from theory to algorithms: an overview." *Journal of physics: conference series.* Vol. 1142. No. 1. IOP Publishing.

Andor, Mark A., and Katja M. Fels. 2018. “Behavioral Economics and Energy Conservation – a Systematic Review of Non-Price Interventions and Their Causal Effects.” *Ecological Economics* 148 (June): 178–210. https://doi.org/10.1016/j.ecolecon.2018.01.018.

Andor, Mark A, Andreas Gerster, Jörg Peters, and Christoph M Schmidt. 2020. “Social Norms and Energy Conservation Beyond the US.” *Journal of Environmental Economics and Management* 103: 102351.

Andor, M. A., Gerster, A., & Peters, J. 2022. “Information campaigns for residential energy conservation”. *European Economic Review,* 104094.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics*. Princeton university press.

Asensio, Omar I, and Magali A Delmas. 2015. “Nonprice Incentives and Energy Conservation.” *Proceedings of the National Academy of Sciences* 112 (6): E510–15.

Athey, S., & Imbens, G. W. (2019). “Machine learning methods that economists should know about.” *Annual Review of Economics*, 11, 685-725.

Athey, Susan. 2017. “Beyond Prediction: Using Big Data for Policy Problems.” *Science* 355 (6324): 483–85.

Athey, Susan, and Guido W Imbens. 2015. “Machine Learning Methods for Estimating Heterogeneous Causal Effects.” *Stat* 1050 (5): 1–26.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. “Generalized Random Forests.” *The Annals of Statistics* 47 (2): 1148–78.

Attari, Shahzeen Z., Michael L. DeKay, Cliff I. Davidson, and Wändi Bruine de Bruin. 2010. “Public Perceptions of Energy Consumption and Savings.” *Proceedings of the National Academy of Sciences* 107 (37): 16054–59. https://doi.org/10.1073/pnas.1001509107.

Attari, Shahzeen Z., Gautam Gowrisankaran, Troy Simpson, and Sabine M. Marx. 2014. “Does Information Feedback from in-Home Devices Reduce Electricity Use? Evidence from a Field Experiment.” {SSRN} Scholarly Paper ID 2545187. Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2545187.

Bahmanyar, Alireza, Abouzar Estebsari, and Damien Ernst. 2020. “The Impact of Different COVID-19 Containment Measures on Electricity Consumption in Europe.” *Energy Research & Social Science* 68: 101683.

Banerjee, Abhijit, Arun G Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O Jackson, Harini Kannan, et al. 2021. “Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization.” National Bureau of Economic Research.

Battocchi, Keith, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. 2019. "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation." https://github.com/microsoft/EconML.

Bonan, Jacopo, Cristina Cattaneo, Giovanna d'Adda, and Massimo Tavoni. 2020. "The Interaction of Descriptive and Injunctive Social Norms in Promoting Energy Conservation." *Nature Energy* 5 (11): 900–909.

Brian C. Prest, Casey J. Wichman, and Karen Palmer. 2021. "RCTs Against the Machine: Experimental Treatment Prediction Methods Recover Can Machine Learning Effects." *Resources for the Future.*

Brülisauer, Marcel, Lorenz Goette, Zhengyi Jiang, Jan Schmitz, and Renate Schubert. 2020. "Appliance-Specific Feedback and Social Comparisons: Evidence from a Field Experiment on Energy Conservation." *Energy Policy* 145: 111742.

Burkhardt, Jesse, Kenneth Gillingham, and Praveen K Kopalle. 2019. "Experimental Evidence on the Effect of Information and Pricing on Residential Electricity Consumption." National Bureau of Economic Research.

Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram. 2020. "Machine Learning from Schools about Energy Efficiency." *Journal of the Association of Environmental and Resource Economists* 7 (6): 1181–1217.

Chen, Huigang, Totte Harinen, Jeong-Yoon Lee, Mike Yung, and Zhenyu Zhao. 2020. "CausalML: Python Package for Causal Machine Learning." https://arxiv.org/abs/2002.11631.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. "Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, with an Application to Immunization in India." National Bureau of Economic Research.

"Commission Recommendation (EU) 2019/1658 of 25 September 2019 on Transposing the Energy Savings Obligations Under the Energy Efficiency Directive, OJ L 275." 28.10.2019. http://data.europa.eu/eli/reco/2019/1658/oj.

Costa, Dora L, and Matthew E Kahn. 2013. "Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment." *Journal of the European Economic Association* 11 (3): 680–702.

Cuerdo-Vilches, Teresa, Miguel Ángel Navas-Martín, and Ignacio Oteiza. 2021. "Behavior Patterns, Energy Consumption and Comfort During COVID-19 Lockdown Related to Home Features, Socioeconomic Factors and Energy Poverty in Madrid." *Sustainability* 13 (11): 5949.

"Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on Energy Efficiency, Amending Directives 2009/125/EC and 2010/30/EU and Repealing Directives 2004/8/EC and 2006/32/EC, OJ l 315." 14.11.2012. http://data.europa.eu/eli/dir/2012/27/2021-01-01.

"Directive (EU) 2018/2002 of the European Parliament and of the Council of 11 December 2018 Amending Directive 2012/27/EU on Energy Efficiency, OJ l 328." 21.12.2018. http://data.europa.eu/eli/dir/2018/2002/oj.

Fezzi, Carlo, and Valeria Fanghella. 2020. "Real-Time Estimation of the Short-Run Impact of COVID-19 on Economic Activity Using Electricity Market Data." *Environmental and Resource Economics* 76 (4): 885–900.

Fontana, Matteo, Massimo Tavoni, and Simone Vantini. 2019. "Functional Data Analysis of High-Frequency Load Curves Reveals Drivers of Residential Electricity Consumption." *PLOS ONE* 14 (6): e0218702. https://doi.org/10.1371/journal.pone.0218702.

García, Sebastián, Antonio Parejo, Enrique Personal, Juan Ignacio Guerrero, Félix Biscarri, and Carlos León. 2021. "A Retrospective Analysis of the Impact of the COVID-19 Restrictions on Energy Consumption at a Disaggregated Level." *Applied Energy* 287: 116547.

Gong, Feixiang, Ninghui Han, Dezhi Li, and Shiming Tian. 2020. "Trend Analysis of Building Power Consumption Based on Prophet Algorithm." In *2020 Asia Energy and Electrical Engineering Symposium (AEEES)*, 1002–6. IEEE.

Granella, Francesco, Lara Aleluia Reis, Valentina Bosetti, and Massimo Tavoni. 2021. "COVID-19 Lockdown Only Partially Alleviates Health Impacts of Air Pollution in Northern Italy." *Environmental Research Letters* 16 (3): 035012.

Henry, Marisa L, Paul J Ferraro, and Andreas Kontoleon. 2019. "The Behavioural Effect of Electronic Home Energy Reports: Evidence from a Randomised Field Trial in the United States." *Energy Policy* 132: 1256–61.

Herzen, Julien, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, et al. 2021. "Darts: User-Friendly Modern Machine Learning for Time Series." https://arxiv.org/abs/2110.03224.

Hong, Tao, and Shu Fan. 2016. "Probabilistic Electric Load Forecasting: A Tutorial Review." *International Journal of Forecasting* 32 (3): 914–38.

Hook, Andrew, et al. 2020. "A systematic review of the energy and climate impacts of teleworking." *Environmental Research Letters* 15.9: 093003.

Houde, Sebastien, Annika Todd, Anant Sudarshan, June A. Flora, and K. Carrie Armel. 2013. "Real-Time Feedback and Electricity Consumption: A Field Experiment Assessing the Potential for Savings and Persistence." *The Energy Journal* 34 (1). https://ideas.repec.org/a/aen/journl/ej34-1-04.html.

Imai, Kosuke, and In Song Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29 (3): 405–15.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2018. "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." *American Economic Journal: Economic Policy* 10 (1): 240–67. https://doi.org/10.1257/pol.20160093.

Jachimowicz, Jon M, Oliver P Hauser, Julia D O'Brien, Erin Sherman, and Adam D Galinsky. 2018. "The Critical Role of Second-Order Normative Beliefs in Predicting Energy Conservation." *Nature Human Behaviour* 2 (10): 757–64.

Jessoe, Katrina, and David Rapson. 2014. "Knowledge Is (Less) Power: Experimental Evidence from Residential Energy Use." *The American Economic Review* 104 (4): 1417–38. https://doi.org/10.1257/aer.104.4.1417.

JK, Aronson, Bankhead C, and Nunan D. 2018. "Confounding, in Catalogue of Biases." *Catalogue of Bias Collaboration*. www.catalogueofbiases.org/biases/confounding.

Kandul, Serhiy, Ghislaine Lang, and Bruno Lanz. 2020. "Social Comparison and Energy Conservation in a Collective Action Context: A Field Experiment." *Economics Letters* 188: 108947.

Karlin, Beth, Joanne F. Zinger, and Rebecca Ford. 2015. "The Effects of Feedback on Energy Conservation: A Meta-Analysis." *Psychological Bulletin* 141 (6): 1205–27. https://doi.org/10.1037/a0039650.

Knittel, Christopher R, and Samuel Stolper. 2019. "Using Machine Learning to Target Treatment: The Case of Household Energy Use." National Bureau of Economic Research.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116 (10): 4156–65.

Kwac, Jungsuk, J. Flora, and R. Rajagopal. 2014. "Household Energy Consumption Segmentation Using Hourly Data." *IEEE Transactions on Smart Grid* 5 (1): 420–30. https://doi.org/10.1109/TSG.2013.2278477.

Li, Chen, Zhenyu Chen, Jinbo Liu, Dapeng Li, Xingyu Gao, Fangchun Di, Lixin Li, and Xiaohui Ji. 2019. "Power Load Forecasting Based on the Combined Model of LSTM and XGBoost." In *Proceedings of the 2019 the International Conference on Pattern Recognition and Artificial Intelligence*, 46–51.

Li, Lechen, Christoph J Meinrenken, Vijay Modi, and Patricia J Culligan. 2021. "Impacts of COVID-19 Related Stay-at-Home Restrictions on Residential Electricity Use and Implications for Future Grid Stability." *Energy and Buildings* 251: 111330.

Li, Richard, Ashwin Shinde, An Liu, Scott Glaser, Yung Lyou, Bertram Yuh, Jeffrey Wong, and Arya Amini. 2020. "Machine Learning–Based Interpretation and Visualization of Nonlinear Interactions in Prostate Cancer Survival." *JCO Clinical Cancer Informatics* 4: 637–46.

Liebe, Ulf, Jennifer Gewinner, and Andreas Diekmann. 2021. "Large and Persistent Effects of Green Energy Defaults in the Household and Business Sectors." *Nature Human Behaviour* 5 (5): 576–85.

Lim, Bryan, Sercan Ö Arik, Nicolas Loeff, and Tomas Pfister. 2021. "Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting." *International Journal of Forecasting*.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. "Explainable Ai: A Review of Machine Learning Interpretability Methods." *Entropy* 23 (1): 18.

Lloyd, Stuart P. "Least squares quantization in PCM." *Information Theory, IEEE Transactions* on 28.2 (1982): 129-137.

Long, Christopher, Bradford F Mills, and Joachim Schleich. 2018. "Characteristics or Culture? Determinants of Household Energy Use Behavior in Germany and the USA." *Energy Efficiency* 11 (3): 777–98.

Lundberg, Scott M, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. "From Local Explanations to Global Understanding with Explainable AI for Trees." *Nature Machine Intelligence* 2 (1): 56–67.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77.

Lynham, John, Kohei Nitta, Tatsuyoshi Saijo, and Nori Tarui. 2016. "Why Does Real-Time Information Reduce Energy Consumption?" *Energy Economics* 54 (February): 173–81. https://doi.org/10.1016/j.eneco.2015.11.007.

Manson, JoAnn E, Chrisandra L Shufelt, and James M Robins. 2016. "The Potential for Postrandomization Confounding in Randomized Clinical Trials." *JAMA* 315 (21): 2273–74.

Marangoni, Giacomo, M. Kochański, I. Malvestio, and K. Korczak. 2022. "Statistical Learning of Residential Electricity Consumption Data, Deliverable D5.1 of the newTRENDs Project, Funded by the H2020 Programme (No. 893311)." https://newtrends2020.eu.

McKerracher, Colin, and Jacopo Torriti. 2013. "Energy Consumption Feedback in Perspective: Integrating Australian Data to Meta-Analyses on in-Home Displays." *Energy Efficiency* 6 (2): 387–405. https://doi.org/10.1007/s12053-012-9169-3.

Menneer, Tamaryn, Zening Qi, Timothy Taylor, Cheryl Paterson, Gengyang Tu, Lewis R Elliott, Karyn Morrissey, and Markus Mueller. 2021. "Changes in Domestic Energy and Water Usage During the UK Covid-19 Lockdown Using High-Resolution Temporal Data." *International Journal of Environmental Research and Public Health* 18 (13): 6818.

Mickwitz, Per. 2003. "A Framework for Evaluating Environmental Policy Instruments: Context and Key Concepts." *Evaluation* 9 (4): 415–36.

Miller, Tim. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Myers, Erica, and Mateus Souza. 2020. "Social Comparison Nudges Without Monetary Incentives: Evidence from Home Energy Reports." *Journal of Environmental Economics and Management* 101: 102315.

Narajewski, Michał, and Florian Ziel. 2020. "Changes in Electricity Demand Pattern in Europe Due to COVID-19 Shutdowns." In *IAEE Energy Forum.(special Issue)*, 44–47.

O'Brien, William, and Fereshteh Yazdani Aliabadi. 2020. "Does Telecommuting Save Energy? A Critical Review of Quantitative Studies and Their Research Methods." *Energy and Buildings*, 110298.

Oreshkin, Boris N, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2019. "N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting." *arXiv Preprint arXiv:1905.10437*.

Panigrahi, Arun. 2020. "Forecasting Residential Electricity Load Demand Using Machine Learning." PhD thesis, Dublin, National College of Ireland.

Parizad, A, and CJ Hatziadoniu. 2021. "Using Prophet Algorithm for Pattern Recognition and Short Term Forecasting of Load Demand Based on Seasonality and Exogenous Features." In *2020 52nd North American Power Symposium (NAPS)*, 1–6. IEEE.

Rojat, Thomas, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. 2021. "Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey." *arXiv Preprint arXiv:2104.00950*.

Samuel, Arthur L. 1967 "Some studies in machine learning using the game of checkers. II—Recent progress." *IBM Journal of research and development* 11.6: 601-617.

Schleich, Joachim, Marian Klobasa, Sebastian Gölz, and Marc Brunner. 2013. “Effects of Feedback on Residential Electricity Demand—Findings from a Field Trial in Austria.” *Energy Policy* 61 (October): 1097–1106. https://doi.org/10.1016/j.enpol.2013.05.012.

Schultz, P Wesley, Jessica M Nolan, Robert B Cialdini, Noah J Goldstein, and Vladas Griskevicius. 2007. “The Constructive, Destructive, and Reconstructive Power of Social Norms.” *Psychological Science* 18 (5): 429–34.

Shapi, Mel Keytingan M, Nor Azuana Ramli, and Lilik J Awalin. 2021. “Energy Consumption Prediction by Using Machine Learning for Smart Building: Case Study in Malaysia.” *Developments in the Built Environment* 5: 100037.

Soava, Georgeta, Anca Mehedintu, Mihaela Sterpu, and Eugenia Grecu. 2021. “The Impact of the COVID-19 Pandemic on Electricity Consumption and Economic Growth in Romania.” *Energies* 14 (9): 2394.

Souza, Mateus. 2019. “Predictive Counterfactuals for Treatment Effect Heterogeneity in Event Studies with Staggered Adoption.” *Available at SSRN 3484635*.

Staffell, Iain, and Stefan Pfenninger. 2018. “The Increasing Impact of Weather on Electricity Supply and Demand.” *Energy* 145: 65–78.

Stewart, James, Annika Todd, and Charles Kurnik. 2020. “Residential Behavior Evaluation Protocol. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, September 2011-August 2020.” National Renewable Energy Lab.(NREL), Golden, CO (United States).

Sunstein, Cass R. 2021. “The Distributional Effects of Nudges.” *Nature Human Behaviour*, 1–2.

Tenailleau, Quentin M., et al. 2021. “Assessing the impact of telework enhancing policies for reducing car emissions: Exploring calculation methods for data-missing urban areas–Example of a medium-sized European city (Besançon, France).” *Urban Climate 38*: 100876.

Thorndike, Robert L. 1953 “Who belongs in the family.” *Psychometrika*.

Tiefenbeck, Verena, Lorenz Goette, Kathrin Degen, Vojkan Tasic, Elgar Fleisch, Rafael Lalive, and Thorsten Staake. 2016. “Overcoming Salience Bias: How Real-Time Feedback Fosters Resource Conservation.” *Management Science* 64 (3): 1458–76. https://doi.org/10.1287/mnsc.2016.2646.

Varian, Hal R. 2014. “Big Data: New Tricks for Econometrics.” *Journal of Economic Perspectives* 28 (2): 3–28. https://doi.org/10.1257/jep.28.2.3.

Varian, Hal R. 2016. “Causal Inference in Economics and Marketing.” *Proceedings of the National Academy of Sciences* 113 (27): 7310–15. https://doi.org/10.1073/pnas.1510479113.

Wang, Dong, Sven Thunéll, Ulrika Lindberg, Lili Jiang, Johan Trygg, and Mats Tysklind. 2022. "Towards Better Process Management in Wastewater Treatment Plants: Process Analytics Based on SHAP Values for Tree-Based Machine Learning Methods." *Journal of Environmental Management* 301: 113941.

Wang, Yi, Qixin Chen, Chongqing Kang, Mingming Zhang, Ke Wang, and Yun Zhao. 2015. "Load Profiling and Its Application to Demand Response: A Review." *Tsinghua Science and Technology* 20 (2): 117–29. http://www.researchgate.net/profile/Yi_Wang137/publication/275365433_Load_profiling_and_its_application_to_demand_response_A_review/links/553e58fb0cf20184050eb255.pdf.

Wastensteiner, Jacqueline, Tobias M Weiss, Felix Haag, and Konstantin Hopf. 2021. "Explainable AI for Tailored Electricity Consumption Feedback–an Experimental Evaluation of Visualizations."

Yukseltan, Ergun, Ahmet Yucekaya, and Ayse Humeyra Bilge. 2020. "Hourly Electricity Demand Forecasting Using Fourier Analysis with Feedback." *Energy Strategy Reviews* 31: 100524.

Zhang, Yingheng, Haojie Li, and Gang Ren. 2022. "Estimating Heterogeneous Treatment Effects in Road Safety Analysis Using Generalized Random Forests." *Accident Analysis & Prevention* 165: 106507.

Imprint

Citation U11f:

Marangoni, G., Kochański, M., Malvestio, I., Korczak, K. (2022): Impact assessment of household-level behavioral interventions via smart-meter data (newTRENDs - Deliverable No. D4.2). POLIMI, Milano.

Institutes:

Fraunhofer Institute for Systems and Innovation Research ISI (Fraunhofer); E3-Modelling (E3M); Technische Universität Wien (TUW); TEP Energy GmbH (TEP); Politenico di Milano (POLIMI); Research and Innovation Centre Pro-Akademia (RIC); WiseEuropa – Fundacja Warszawski Instytut Studiów Ekonomicznych i Europejskich (Wise); Zentrum für Energiewirtschaft und Umwelt (e-think)

Authors

giacomo.marangoni@polimi.it

maksymilian.kochanski@proakademia.eu

irene.malvestio@polimi.it

katarzyna.korczak@proakademia.eu

Date of release

04/2022

This project has received funding from the European Union´s Horizon 2020 research and innovation programme under grant agreement no. 893311.