Evaluating price-based demand response in practice — with application to the EcoGrid EU Experiment

Guillaume Le Ray, Emil M. Larsen, Pierre Pinson, *Senior Member, IEEE*

***Abstract*—Increased emphasis is placed today on various types of demand response, motivated by the integration of renewable energy generation and efficiency improvements in electricity markets. Some advocated for the development of price-based approaches, where the conditional dynamic elasticity of final users is exploited in the power system, e.g. for system balancing. However, very few real-world experiments have been carried out and price-based demand response has consistently been found difficult to assess and quantify. It is our aim here to describe an approach to do so, as motivated by the large-scale EcoGrid EU experiment. In this project, 1900 houses were equipped with smart meters and other automation devices in order to adapt consumption to real-time electricity prices every five minutes, while monitoring it with the same resolution. Our approach first relies on the clustering of residential load observations that behave similarly within a given experiment. Then, a clinical testing approach, based on a test and a control group, is adapted to assess whether price-responsive loads were actually responsive or not. Interestingly, in the deployment phase of the project, the results show that houses could be deemed price-responsive on some test days, while results were inconclusive on some others.**

***Index Terms*—Clustering, demand response, electric load modelling, smart grid, time-series analysis.**

I. Introduction

TARGETS TO increase the proportion of renewable energy production to 27% by 2030 across all 28 EU member states [1] present significant technical challenges, since existing markets, services and technologies are unlikely to be robust enough to cope with the expansion of variable power generation, also with limited predictability. Among the various options to support large-scale renewables penetration like wind and solar power, Demand Response (DR) has emerged as a popular approach, with its natural advantages and caveats [2]. Applications of DR to provide system services are numerous, e.g., frequency control [3], congestion management [4], distribution grid services [5] and overall system balancing [6]. Some also discuss long term impacts on grid planning [6]. The Ecogrid EU project placed emphasis on DR for overall system balancing, where imbalances mainly originate from wind power forecasting errors. Recent developments in that direction follow the concepts of *(i)* direct control, where a higher-level operator would somehow operate these electric loads, and *(ii)* control by price, where advantage is taken of the elasticity and cross-elasticity of electric power consumers. There obviously are obstacles in rolling out DR, including the non-flexibility of demand [7] and the low participation due to information asymmetries [8]. Control by price has additional difficulties over direct control due to the complexity in predicting response to price variations [9], although forecasting models and control schemes that make effective use of them have been researched [10].

From articles presenting early DR programmes [11] to those reviewing DR state-state-of-the-art deployment [12]–[15], some highlight problems of communication between the different units at or after the deployment process, resulting in non-responsiveness of the controllers. However techniques that can readily identify whether a set of electric loads is price-responsive remain lacking, while this may be crucial in practice. This issue is of particular relevance during the deployment phase of demand response equipment and programs. Indeed a logical subsequent step after deploying necessary hardware and software is to control that the different elements communicate as expected, react to the right information, or simply to verify that the overall concept functions.

The present paper introduces a proposal test-control method to assess whether or not electric loads are price-responsive or not. The principle of comparing control and test groups has been extensively used in the medical industry to evaluate the efficiency of a treatment for over 200 years [16], and more recently in the electricity field, industrials working on load research practices have been using this approach to develop Customer Base Line (CBL) and evaluate candidate customers under DR [17]. This method has the advantage of having both the candidate customers and CBL to be exposed to the same weather conditions. Such an approach aims at assessing through hypothesis testing whether loads are responsive or not, which is a basic question to answer before aiming for a quantification and characterization of that response.

Prior to undergoing this test-control analysis, electrical loads are clustered based on similar behavior within a given experiment (i.e., a test day with a given price profile). This allows to identify electric loads that do not respond as expected, while sorting subgroups of responsive households. Note that here the terms 'household' are used with the same meaning as 'load' and do not consider the consumer behavior as the response is automated by a controller; when we write 'responsive loads' or 'responsive households' the reader should read 'responsive controllers which have modified significantly the electricity consumption'. The value of the clustering step of our methodology also lies in the dimension reduction of the problem since, instead of trying to assess whether each and every household in a large-scale demand response experiment

G. Le Ray, E.M. Larsen and P. Pinson are with the Centre for Electric Power and Energy, Technical University of Denmark, Kgs. Lyngby, Denmark (email: {gleray,emlar,ppin}@elektro.dtu.dk).

This work was partly supported by the European Commission through the project EcoGrid EU (grant ENER/FP7/268199).

Manuscript submitted ...

(with 1000 households or more) is responsive or not, a fully data-driven clustering step narrows down the analysis by focusing on a low number of subgroups of households with similar dynamic characteristics. This may also be seen as having the side benefit of pinpointing electric loads that could be useful in providing specific grid services such as balancing and congestion management, in view of the characteristics of their response.

Existing literature related to clustering applications (also referred to as segmentation) focuses on profiling, to group the consumers with similar energy consumption patterns [18], [19], or on modelling, to obtain more homogeneous data to improve forecasting accuracy [20], [21]. However, similar approaches using clustering to exclude electric loads that are not responding to the price have not been found in the literature, despite interest from industry in knowing whether a smart controller is responsive or not [17].

The development of this methodology was originally motivated by, and then applied to, the EcoGrid EU demand response experiment, in which 1900 houses and 100 industrial loads receive new electricity prices every five minutes [22]. On the Danish island of Bornholm where the experiment takes place, the majority of the participants have resistive electric heating and heat pumps installed. Their controllability, combined with the heat capacity of the buildings, yields virtual electric power storage. Houses are equipped with smart meters reporting consumption in real-time, as well as a range of automated controllers that make provision of DR convenient by enabling controllability of a wide range of small-scale Distributed Energy Resources (DERs) in a cost-efficient manner. The automated controllers are proprietary and were developed by different companies. In this study, they are therefore considered as black boxes. However, it is known that these rely on state-of-the-art control techniques used for DR, like hysteresis control and economic model predictive control, allowing to schedule consumption optimally considering weather and price forecasts, as well as customer preferences in terms of comfort.

The prices seen by these electric loads originate from the EcoGrid EU market. It was primarily designed to support balancing when larger shares of renewables are present in the power system, yielding additional and more variable balancing needs. In EcoGrid EU, knowledge of the power system state is updated every five minutes. This higher temporal resolution, compared to the hourly time units broadly used in deregulated power systems today, naturally allow to better adapt to dynamic balancing needs. Another key aspect of the market is that it is bidless for demand, hence reducing risk and increasing convenience for small customers who would not otherwise participate. A full introduction to the market behind price generation in the EcoGrid EU experiment is given in [23]. The first phase of the EcoGrid EU project was completed in early 2014, where price-responsive controllers from two different manufacturers were installed in 1200 houses. The price-responsiveness of participants was analysed and eventually validated using the clustering and test-control methods presented here.

The paper is structured as following. Section II presents the empirical framework of the experiment, with particular emphasis on the data and various test-cases to be analyzed. Our methodology is described in Section III, by first introducing the clustering approach for identifying fully non-responsive households and subgroups of responsive electric loads, followed by the test-control method to assess whether these responsiveness can be seen as genuine price-responsiveness. The results for the roll-out phase of the EcoGrid EU experiment are used as an illustration in Section IV. The paper ends with conclusions and perspectives for future work in Section V.

II. Empirical Framework

The analysis is data-driven and the processes of generation of the data used in this paper are described in the following section. Details on how the controllers operate were kept confidential by the manufacturers.

A. Data Presentation

The datasets consist of electricity consumption for each candidate household with a resolution of 5 minutes. Real-time price series have the same temporal resolution, allowing for the joint analysis of the dynamics of both price and consumption series. Only consumption related to space heating varies as a function of prices based on the controllers deployed for heat pumps and resistive electric heating.

B. Customer Base Line Generation

Throughout the initial phase of the demonstration, households were recruited and then made price-responsive gradually. Some households had their automation disabled deliberately by the central operator, while others had their automation disabled due to reported technical problems. These households were gathered and averaged to form the CBL. Due to the random nature of technical problems, the composition of the test and CBL groups varied from one test-case to the next. Test and CBL groups also varied according to the number of households using one of two control-equipment types and according to different heating types (heat pump or resistive electric heating). As the size of the CBL and participant groups differ throughout the overall experiment, this influences the resulting data analysis and especially the estimated confidence intervals and hypothesis tests performed.

C. Test Cases Presentation

In order to test the controllers, test cases were designed to stress and assess their price-responsiveness with extreme price variations. As the energy consumption should be a function of the price, a significant change in the electricity consumption is expected when such extreme price variations occur [24]. More precisely, a variation in price is to be seen as an incentive for modification of electricity consumption: upwards when the price goes down, and downwards when the price goes up. Table I and Fig. 1 gives a summary of the price variations applied during each test case. All the test cases have the same duration of 24 hours. The negative prices are strong incentives to consume electricity send to controllers as they are making

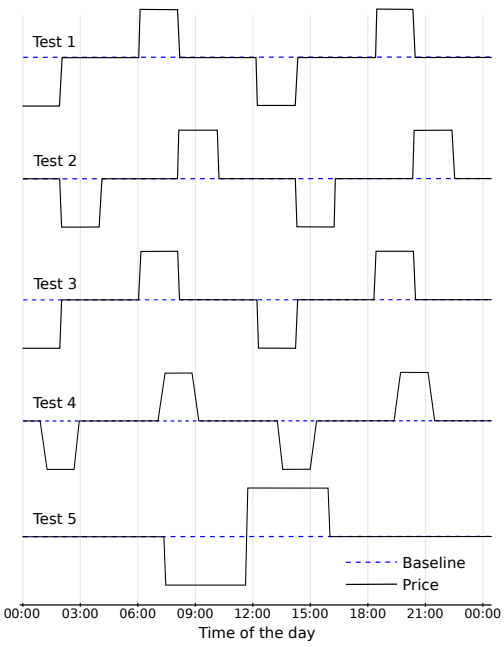

Fig. 1: Price signals broadcast during the different test cases .

money by doing so. In contrary high positive prices are design to decrease immediately the consumption.

TABLE I: PRICE VARIATIONS DURING THE TEST-CASES [24].

Test period	Min / Max (€/MWh)	Baseline (€/MWh)	Test signal
25/10/2013	-53.7 / 148.0	47.2	Test 1
07/11/2013	-53.7 / 148.0	47.2	Test 2
21/11/2013	-61.3 / 140.3	39.4	Test 3
27/11/2013	37.4 / 41.4	39.4	Test 3
06/12/2013	-134.5 / 134.5	0	Test 5
10/12/2013	-134.5 / 134.5	0	Test 5
11/12/2013	-134.5 / 134.5	0	Test 5
12/12/2013	-134.5 / 134.5	0	Test 5
20/01/2014	-134.5 / 134.5	0	Test 5
21/01/2014	-134.5 / 134.5	0	Test 5
22/01/2014	-134.5 / 134.5	0	Test 5
23/01/2014	-134.5 / 134.5	0	Test 5
08/03/2014	-61.3 / 140.3	39.4	Test 3
11/03/2014	37.4 / 41.4	39.4	Test 3
09/04/2014	-61.3 / 140.3	39.4	Test 4
13/04/2014	-61.3 / 140.3	39.4	Test 4

III. METHODOLOGY FOR ASSESSING PRICE-RESPONSIVE BEHAVIOR

A. A Non-supervised Classification for Dimension Reduction and to Identify Sub-groups

A natural way to reduce dimension and to extract information from a large and noisy dataset is to group it into more homogeneous clusters. Each of these clusters exhibit more homogeneous characteristics of its individuals than the overall dataset does [25]. Consequently, clustering can be used to exclude groups which could be considered as outliers [17]. In addition, it emphasizes characteristic patterns in consumption, which may implicitly include the consumption variations due to changes in price.

As it is most likely the case for any real-world experiment, it was observed within the EcoGrid EU demonstration that uncertainty existed in the actual price-responsiveness of heat appliance controllers during DR experiments. This may be due to customers being able to interact with controllers - turning them off or changing comfort settings. Other issues, e.g., bad choice of location for temperature sensors used by controllers, can also result in households not being responsive (or just a little) at certain times. A number of other punctual technical problems can affect the responsiveness of these heat appliance controllers. Therefore, employing clustering for identifying and isolating these outliers can focus our analysis on the DR of well-functioning installations. On a more practical level it generates a list of targets to troubleshoot for the technicians. The time period (24 hours), the replications (16 tests) in the experiment, the amplitude of the incentives and consequently the amplitude of the responses leave little doubt that the largest part of the consumption variation is from variations in the controllers and not from changes in consumers preferences.

Clustering approaches have been extensively described in the literature. The interested reader is for instance referred to [26] for an overview of clustering algorithms and [25] for applications in electric load analysis. Out of this wealth of algorithms, the most suitable one to be used depends upon the data setup and our a priori knowledge of the expected output (e.g., the number of clusters to be obtained) [27]. Hierarchical clustering permits to effectively choose the number of clusters, a posteriori, according to the so-called dendrogram, which is a clustering tree where the level of details (and the number of clusters) is increasing as its branches are further divided. An example dendrogram used to cluster 35 households in one of the EcoGrid EU experiment is shown in Fig. 2. Hierarchical clustering is a non-supervised classification method where individuals are grouped according to their relative distances in a similarity space determined by a set of variables [28]. Hierarchical clustering can be performed in an agglomerative or divisive manner. The former approach starts with each household as a cluster and ends up with one cluster (bottom-up approach), while the latter one sees the whole set of households as one cluster to start with and eventually ending with each household as a cluster (top-down approach). Their outputs are similar, but Hierarchical Agglomerative Clustering (HAC) is known to be faster to compute.

Here our households may naturally have different average consumption levels depending on the house types, number of inhabitants and human behavior. Consequently, some form of alignment is needed to make them all comparable in order to measure some kind of distances between them. However the variance σ_i^2 of the time-series from each household i should not be affected, as the variability in amplitude of the adjustment in consumption during DR events is of high importance. For each and every test case in the experiment, electric power consumption series were centered on their average consumption, by subtracting the mean consumption on a per-household basis, over the entire test case. Considering the original power consumption series $\mathbf{x}'_i = \{x'_{i,1}, \ldots, x'_{i,t}, \ldots, x'_{i,T}\}$

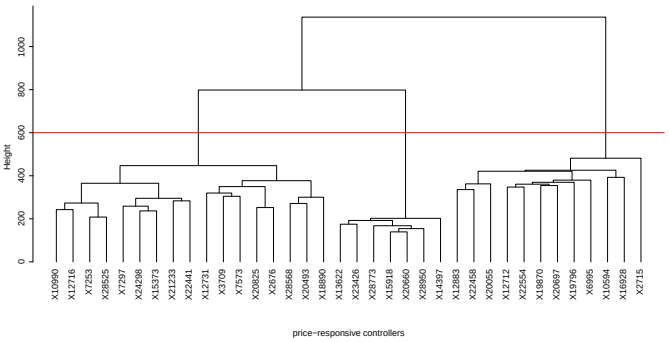

Fig. 2: dendrogram for the clustering of of 35 households in one of the EcoGrid EU experiment. The red line indicates the cut to be made to obtain 3 clusters.

for household i ($i = 1, \ldots, I$), with t the time index, this reads

$$x_{i,t} = x'_{i,t} - \frac{1}{T}\sum_{t=1}^{T} x'_{i,t}, \quad i = 1, \ldots, I, \ t = 1, \ldots, T \tag{1}$$

$\mathbf{x}_i = \{x_{i,1}, \ldots, x_{i,t}, \ldots, x_{i,T}\}$ is the resulting centered power consumption series for household i, with I the number of households at time t. The series $\mathbf{x}_i$ has the same dynamics and amplitude as $\mathbf{x}'_i$, though centered on 0, thus allowing to better compare the higher-order dynamics of the various households [29], [30].

In our experimental framework, the hypothesis is that if a household is active and receives a price variation during a DR event, the consumption should be affected. The variation in consumption is not expected the same for all houses because of their prior status (e.g. temperature, controller setup), nevertheless it should be possible to cluster similar patterns of consumptions' variation as they are expected to react. In that context, the chosen distance for the clustering approach ought to account for covariances between the consumption series. In our experimental framework, the space we have to explore has the dimension of the number of measurements performed in time. With a temporal resolution of 5 minutes and a test case duration typically of 24 hours, this translates to fairly large dimensions. However, it is expected that power consumption observations are serially correlated, i.e., not independent from one time instant to others. In other words, the effective dimension of the space within which the consumption patterns are observed is clearly less than the number of time steps T. The chosen distance for the clustering approach ought to reflect that aspect. The Mahalanobis distance [31], which fulfills this requirement, is then adopted. For two series $\mathbf{x}_i$ and $\mathbf{x}_j$, it is defined as

$$d(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^\top S_{ij}^{-1} (\mathbf{x}_i - \mathbf{x}_j) \tag{2}$$

where S_{ij} is the covariance matrix between the two time-series. However, the covariance matrix S_{ij} may happen to be singular when the number of households (I) is smaller or about the same as the number of data points (T) in the time-series [32]. This problem arises often while working with time-series as the number of data points can be extensive compared to the number of households. To prevent such issues with singularity, S_{ij} is replaced in (2) by a shrunk covariance matrix S^*_{ij}. Shrinkage is an efficient way to obtain a non-singular closest estimate of the original covariance matrix S_{ij}. It is calculated as

$$S^*_{ij} = \lambda T_{ij} + (1 - \lambda) S_{ij} \tag{3}$$

where T_{ij}, commonly referred to as the target, is a diagonal matrix formed with the element on the main diagonal of the original covariance matrix S_{ij} [32]. λ is the shrinkage coefficient. S^*_{ij} is a trade-off between a highly-structured matrix (T_{ij}) and a non-organized one (S_{ij}), while λ allows controlling the balance between the two [33]. We set

$$\lambda = \begin{cases} \lambda^*, & \text{if } \lambda^* \leqslant 1 \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

with

$$\lambda^* = \frac{\sum_{m \neq n} \widehat{\sigma}(c_{mn})}{\sum_{m \neq n} c_{mn}^2} \tag{5}$$

where c_{mn} are the components of the (sample) covariance matrix S_{ij} and $\widehat{\sigma}(c_{mn})$ their estimated variance [34].

HAC is a fairly general framework, given a metric suitable for the data at hand (e.g., the Mahalanobis one used here). Similarly, one may flexibly choose the way to regroup individuals within clusters. The most common one is the Ward's method, also known as minimum-variance method. It aims to minimize the increase of the within-cluster sum of squared distances, E, at each iteration of the agglomerative process [26],

$$E = \sum_{k=1}^{K} \sum_{\mathbf{x}_i \in C_k} d(\mathbf{x}_i, \mathbf{g}_k)^2 \tag{6}$$

where K is the number of clusters, $\mathbf{x}_i \in C_k$ the households in cluster C_k and $\mathbf{g}_k$ the center of gravity of cluster C_k, defined as

$$\mathbf{g}_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i \tag{7}$$

where n_k is the number of households in C_k.

Following [35], the total variance of a set of households, after clustering, can be expressed as the sum of the within-cluster variance plus the between-cluster's center variance. Consequently, since the Ward's method aims at minimizing the increase of within-cluster variance at each iteration, it also maximizes the variance between cluster centers. The resulting clusters can then be seen as the most homogeneous possible subgroups from the set of households. The HAC algorithm is illustrated in Fig. 3, starting with each household being its own cluster. It then iterates until all households are merged into a single cluster. The result of the HAC is conveniently represented in a dendrogram such as that in Fig. 2. The dendrogram is a basis to decide on how many clusters should be chosen. The decision of where to cut the tree depends on the structure of the tree and the goal of clustering. If the goal is to have a clear and precise information on each cluster, a higher number of cluster will be favored. Conversely, if the goal is to isolate outliers, a lower number of clusters will be favored. It is then difficult to implement an automated routine to select the number of clusters. The decision is based on

our expertise in interpreting the structure of the tree and thus subjective [36].

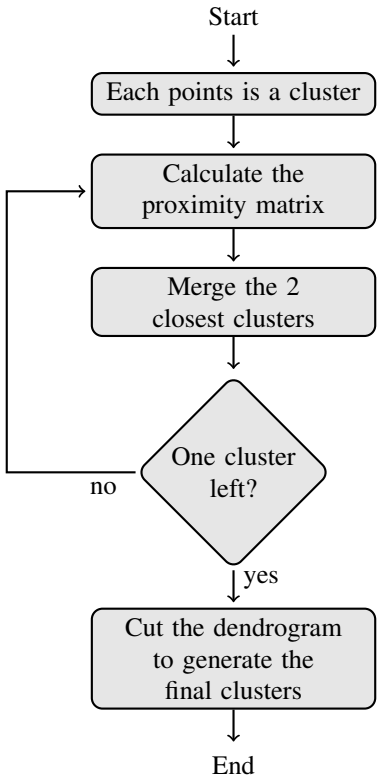

Fig. 3: The Hierarchical Agglomerative Clustering (HAC) algorithm.

After computing the HAC, the information contained in the different clusters should be summarized. When it comes to time-series, the clusters' averaged time-series is a suitable way to represent the specificities of each cluster. One of our test cases, with 5 averaged cluster time-series identified from the dendrogram, is shown in Fig. 4 together with the averaged time-series of the CBL, as well as the corresponding price signal. Such representation allows clusters with reactive adjustment to the price variations (if compared to CBL) to be sorted apart from those that do not adjust during the DR event or show erratic patterns (e.g., due to technical problems). These are consequently not considered in the subsequent analysis. In the example of Fig. 4, the households from the clusters 2 and 5 are to be excluded from the test group, since cluster 2 follows the CBL while cluster 5 has no daily variations which most likely means that the households are empty. As these outliers are removed, the data quality of the treatment group is improved and eases the subsequent qualitative and quantitative analysis. When mentioning test groups in the remainder of the paper, we refer to those subgroups selected after the clustering was performed.

B. A Clinical Trial Test Approach

Clinical trials were historically developed in the pharmaceutical industry. Owing to the variety of potential responses of biological organisms as individuals, it became common to

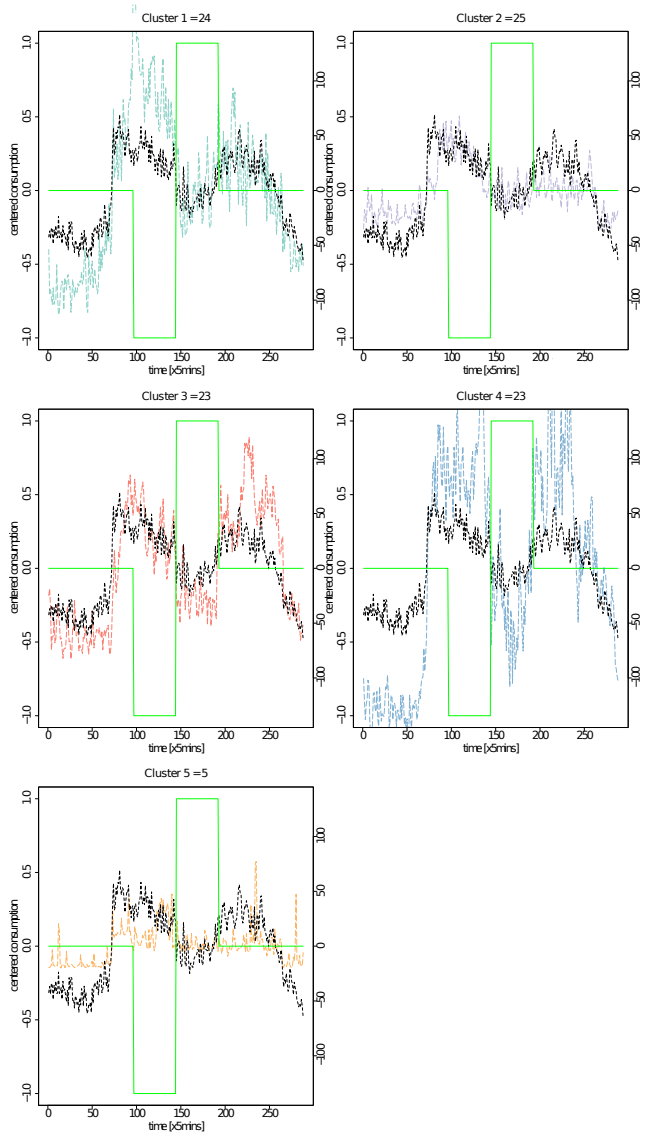

Fig. 4: The averaged time-series is calculated for each cluster in the test group and displayed as a colored dashed line. The black dashed line is the averaged time-series of the CBL and the green line is the price.

perform tests on populations instead, thereby smoothing the potential negative effect of individual features on an overall assessment. In the present case, we can employ a similar clinical trial test approach since our data comes from a reference (CBL) and a test group, while our interest lies in the difference in consumption between these two groups [37]. Moreover, the inherent uncertainty on the responsiveness (response to a treatment) resulting from the absence of homogeneity in the test group as well as in the reference group (e.g., behavior of the user, thermal comfort setup, energetic profile of the buildings), supports the idea that a clinical trial approach is relevant here. The question we aim to answer can be formulated as *Do price variations induce significant changes in power consumption patterns?*

The results of the clustering on households recruited in the DR program from the different groups exposed in Table II, can be analyzed in two different ways. On the one hand, one can

visually assess whether the average response of the selected clusters from each group is responsive by direct comparison with the CBL during the DR event. The purpose of visual inspection is to show that DR works for some clusters, and not for others, to a non-scientific audience who is not familiar with more objective statistical methods. However such an analysis cannot conclude on the significance of the response observed, while relying on expert knowledge at evaluating variation in patterns. On the other hand, this can be tested more rigorously in an hypothesis testing framework (see Section III-C). The purpose of hypothesis testing is to satisfy a scientific audience who requires a degree of objectivity when the results (e.g., lower unit electricity costs for the consumer) are presented.

Fig. 5 is an example of a case used for visual evaluation of a given test case. By observing the dynamics of the mean consumption series of the test group compared to the CBL, one may conclude on the responsiveness of that test group based on confidence intervals. Experience with such consumption data shows that it does not follow a Gaussian distribution. Hence, a nonparametric approach (Non-Studentized pivotal method) is used to obtain confidence intervals. More specifically, we employ a common resampling technique known as bootstrap [38] to generate them. From all the 5000 resampled average time-series, 95% confidence intervals defined by 2.5% and 97.5% quantiles of the distributions of bootstrap samples are obtained.

From visual inspection of Fig. 5, one may infer that the behavior of the test group is different from that of the CBL when the confidence intervals are not overlapping (for example, from 7:05 to 8:05). In other situations, when the confidence intervals overlap or when the average time-series lies within the confidence intervals of another one, one cannot conclude. A more detailed analysis of Fig. 5 shows that the test group exhibits higher consumption during the low price period and lower consumption in the high price period with respect to the CBL. The lower consumption in the period 23:05 to 5:05 is induced by the smart controllers in the experiment shifting load to the lower price period that starts at 07:05. Smart controllers receive a day-ahead price forecast (as well as an hour-ahead price forecast every half hour) allowing them to schedule consumption in an optimal manner. The value of the relative real-time price with respect to recent and limited forecasted prices therefore contributes to visual estimation of whether a test group is price response or not. For example, in Fig. 5, the relative price is high in the period 23:05 to 5:05, so it is expected that a price-responsive cluster would have lower consumption than the CBL during this period.

C. Hypothesis Testing to Assess Price-responsiveness

A standard way to assess results in a clinical trial test is to employ hypothesis testing. The hypothesis obviously depends on the question, e.g., is the test group's consumption different than that of the CBL during a DR event? In this question it can even be specified lower or higher instead of 'different'. Based on this hypothesis, a test is formulated and applied to the data. The method used to analyze the hypothesis test should be chosen according to the assumptions on the sample values' distribution. The aim of the EcoGrid EU DR program is to displace electric power consumption from periods with higher prices to periods with lower prices. Whether this goal is achieved or not can then be determined based on the economic value to the households, i.e., in relation to cost per unit of electricity consumed. Consequently here, hypothesis testing may allow us to objectively state whether a test group is price responsive or not. We use a framework similar to that of conventional clinical trial tests, with a type I error threshold α of 0.05.

A hypothesis test can be formulated, since the average cost of a kWh of electricity consumed during a test-case by the test group should be lower than the cost for the CBL during the same period. The average unit cost $\bar{C}_i$, for a test case with T time steps, is calculated as

$$\bar{C}_i = \frac{\sum_{t=1}^{T} C_{ti} P_t}{\sum_{t=1}^{T} C_{ti}} \tag{8}$$

where C_{ti} is the consumption of electricity from household i at time t and P_t the price at time t. A simple observation of the average unit cost distributions tells us that the variances of the 2 samples are different and that they may have heavy tails. Therefore, standard parametric tests are excluded. The Mann-Whitney test (also known as the Wilcoxon rank sum test) is a convenient solution, since the number of households in each of the subgroups is large. A one-sided Mann-Whitney test is performed on the ranks. The hypotheses are the following,

$$\begin{aligned} H_0 &: \mu_{test} \geqslant \mu_{CBL} \\ H_1 &: \mu_{test} < \mu_{CBL} \end{aligned} \tag{9}$$

where μ corresponds to the sum of ranks, H_1 is the one-sided tailed alternative hypothesis and H_0 is the null hypothesis. The null hypothesis means that the activity of the price-responsive controllers is not significantly modifying the average unit cost, so that it could be considered lower than the control group average unit cost. If the H_0 is rejected, the alternative hypothesis is confirmed statistically.

The one-sided tailed alternative hypothesis is more restrictive than the two-sided tailed standard hypothesis test, as it specifies that the samples should not only be different, but that the test sample's mean should be lower than the control sample's mean. The Mann-Whitney test defines the statistic U with the following formula

$$\begin{aligned} U \quad = \quad & \min\Big(n_1 n_2 + \frac{n_1(n_1-1)}{2} - R_1, \\ & n_1 n_2 + \frac{n_2(n_2-1)}{2} - R_2\Big) \end{aligned} \tag{10}$$

where n_1, n_2 are the size of the 2 samples and R_1, R_2 are the sum of the ranks for these two samples respectively. U follows a normal distribution and we can calculate the p-value as

$$P(U \geqslant U_{1-\alpha} | \mu_{test} \geqslant \mu_{CBL}) \tag{11}$$

The P-value can be seen as the probability of obtaining a test statistic result at least as extreme or as close to the one that was actually observed, assuming that the null hypothesis is true. The test is considered significant when the p-value is lower than the type I error threshold α, which is the chance that we

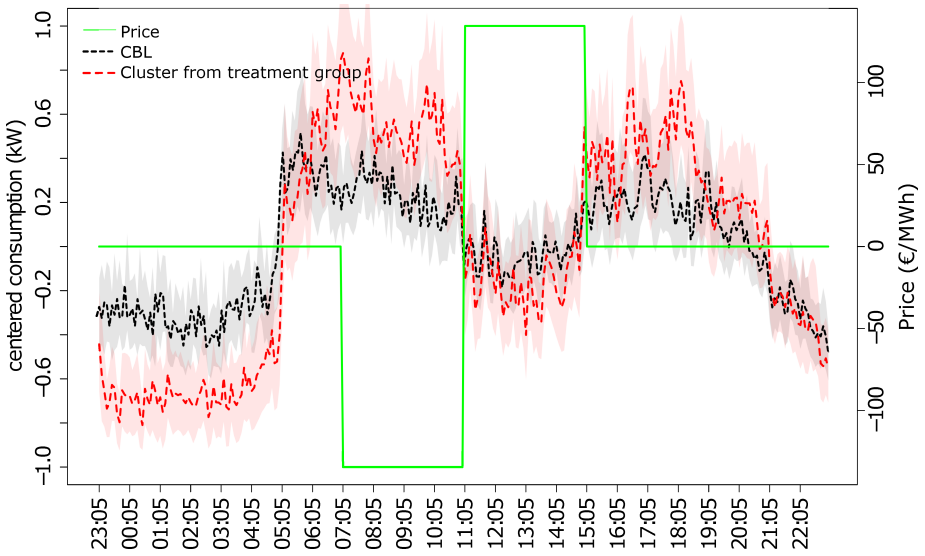

Fig. 5: The average time-series from the CBL and a cluster from the test group with their respective 95% confidence intervals generated from the bootstrap. The green line is the price.

mistakenly reject the null hypothesis (that the samples' means are different). All the details related to statistical aspects can be found in [39].

IV. Results

A. Clustering Results

The cluster analysis aims to identify the price-responsive participants in the test group (possibly in the form of various subgroups) and to separate them from obviously non-responsive households. Emphasis is placed here on how many households are kept in the analysis from the original subgroups after computation of the HAC, as it influences the subsequent analysis.

TABLE II: Number of households in various test cases: number in the test group (number deemed price-responsive after clustering) / number in the CBL.

Date	Manufacturer 1 Electric Heating	Manufacturer 1 Heat Pump	Manufacturer 2 Electric Heating
25/10/2013	68 (48) / 288	36 (24) / 197	88 (55) / 82
07/11/2013	65 (58) / 289	36 (33) / 197	88 (61) / 92
21/11/2013	67 (61) / 292	36 (36) / 200	87 (75) / 94
27/11/2013	66 (55) / 292	36 (34) / 201	89 (78) / 91
06/12/2013	—	—	115 (74) / 99
10/12/2013	—	—	103 (84) / 91
11/12/2013	—	—	100 (70) / 86
12/12/2013	—	—	106 (69) / 89
20/01/2014	—	—	230 (194) / 105
21/01/2014	—	—	223 (121) / 100
22/01/2014	—	—	230 (110) / 104
23/01/2014	—	—	229 (107) / 105
08/03/2014	30 (30) / 324	20 (17) / 236	237 (125) / 75
11/03/2014	38 (38) / 317	24 (18) / 229	232 (171) / 76
09/04/2014	101 (99) / 249	58 (43) / 188	249 (197) / 109
13/04/2014	38 (38) / 311	24 (22) / 222	269 (188) / 114

The clusters are visually selected by comparing the averaged consumption time-series of each cluster to the averaged CBL consumption time-series during event with price variations (Fig. 4). If the averaged consumption time-series of a cluster seems to be flat (no activity, e.g., as for cluster 5 in Fig. 4), following the same pattern as the averaged CBL time-series (e.g., as for cluster 2 in Fig. 4) or showing unexpected pattern, it will be excluded from the dataset used in the evaluation of the price responsiveness. When all the clusters are non price-responsive, only the aberrant ones will be removed.

Table II gives a summary of the clustering selection; the range of the selection from the original data goes from 52% to 100%. In other words, a maximum of a half (48%) of the smart controllers were in the test group, but did not visually appear to be price-responsive. The graphical representation of clusters is also useful for identifying different types of price-responsive behavior. For example, in Fig. 4, cluster 1 gathers the controllers which have been stimulated by the first price variation, while cluster 3 gathers the ones which have been stimulated by the second price variation and cluster 4 gathers the ones which have reacted to both stimuli. It also illustrate the differences of behavior between the manufacturers as the price-response strategies and constraints are implemented differently. Such information was not known beforehand, and brought more insight on how a set of controllers behave at the occasion of large price variations. However, this paper does not focus on this aspect, but it worth mentioning it as it is a good way to illustrate it.

B. Results of the Clinical Trial Test Approach

The chosen clusters are used to generate graphical overviews of each group during the different test-cases (Fig. 5). Table III summarizes the visual evaluation of the graphs displaying the averaged time-series associated with the 95% confidence intervals of the treatment and CBL groups for each manufacturer, equipment type and for different test-days. Results here should be interpreted as, for each experiment, whether it was possible to find one or more clusters that could be seen as price responsive, or not.

In the roll-out phase of the EcoGrid EU demonstration, controllers and other infrastructures were continually developed and improved, which explains the improvement of the DR as the heating period went on.

C. Results of the Hypothesis Testing

The main goal of the EcoGrid EU project is to push electricity consumption during periods of high prices to periods of low electricity prices. This means an economic evaluation can be done, by comparing the average unit cost of selected test groups to the CBL. In this case, hypothesis testing could be applied to each and every identified clusters, or only to those where visual assessment indicated that price response may be present. As for the visual assessment before, the test is applied to all clusters that were not discarded through the clustering analysis, for instance since deemed as outliers or clearly non-responsive.

TABLE III: THE COLOR OF THE CELL RETURN THE RESULTS OF THE VISUAL EVALUATION; GRAY IS RESPONSIVE, LIGHT GRAY IS NON-RESPONSIVE. THE FIGURE IS THE P-VALUE FROM THE MANN-WHITNEY TEST. SIGNIFICANT TEST AT $\alpha = 5\%$ ARE SHOWN IN BOLD AND ITALIC.

Date	Manufacturer 1 Electric Heating	Manufacturer 1 Heat Pump	Manufacturer 2 Electric Heating
25/10/2013	0.98	0.44	0.20
07/11/2013	0.39	***0.0013***	0.88
21/11/2013	0.95	0.20	0.09
27/11/2013	0.34	0.34	0.81
06/12/2013	—	—	0.13
10/12/2013	—	—	***0.00022***
11/12/2013	—	—	***0.0015***
12/12/2013	—	—	0.12
20/01/2014	—	—	0.99
21/01/2014	—	—	0.22
22/01/2014	—	—	0.21
23/01/2014	—	—	0.63
08/03/2014	0.54	0.59	***0.0068***
11/03/2014	0.86	0.80	0.28
09/04/2014	***0.0034***	***0.0096***	0.21
13/04/2014	0.96	0.12	***0.0014***

The Table III shows the Mann-Whitney test's results for the different test periods. A standard type I error threshold is chosen ($\alpha = 5\%$). The significant tests are shown in bold and italic. The comparison between the results from visual evaluation and the hypothesis testing in Table III exposes the difference between price-responsiveness which can be visually noticed but not statistically validated using the measure of unit cost, and the price-responsiveness that does have a significant economical impact on the average unit cost. The results show that towards the end of the roll-out of the EcoGrid EU project, it was possible to visually and rigorously find differences between CBL and test groups (manufacturer 1 electric heating, manufacturer 1 heat pump and manufacturer 2 electric heating), indicating a price-responsive behavior overall. The improvement of the responsiveness over time is a direct consequence of the tunning operated on the controllers during the experiment. Further steps in such an evaluation work would consist in quantifying and characterizing this price-responsiveness, while also assessing if this corresponds to the maximum response that could be provided by these groups of households.

V. CONCLUSIONS

The method presented in this paper shows how a systematic evaluation of DR can be done even with datasets that contain outliers, noise, and other undesirable effects. The clustering can easily be generalized to other time series classification, although scalability to data with more observations remains an area for inquiry. We have successfully applied it to 2 weeks data with a resolution of 5 minutes, but further work should investigate clustering of time-series with more observations. Clustering based on the coefficients of an auto-regressive model of each subject may be viable.

The methodology established provides a springboard to further understand the different types of DR present in residential loads. User interaction with DER controllers is expected to have a large impact on the DR available, and the HAC used to separate useful households from those which do not appear extremely effective in this circumstance.

From a widespread power system perspective, being able to identify which customer segments exhibit a price response is important for grid operators looking to identify and invest in customers to participate in new DR schemes. Such clustering may also be a useful technique to decide additional financial reward for customers who perform best, in the form of a capacity payment, perhaps funded by the same public service obligations (PSOs) that support renewable generation.

Comparing treatment subgroups to the CBL graphically is also useful for presenting the differences in consumption to a broad audience in an intuitive manner. However, visual interpretation is not a statistically valid way of confirming a response. Therefore, the 2-sample Mann-Whitney test comparing the averaged unit cost of price-responsive and non price-responsive subgroups supplements the graphical approach well, as it allows us to validate or reject hypothesis for each test-case. This analysis answers one of the key points of the demonstration: cost can be reduced for some consumers. Obviously, a necessary further step is to characterize and quantify the responsiveness of electric loads. This has been the focus of our further research over a 8 month live experiment in the EcoGrid EU project, which kicked off after the first assessment results presented here allowed to verify the demand response potential in our set of electric loads.

ACKNOWLEDGEMENT

The authors would like to thank all their partners in the EcoGrid EU project for their contribution to the methodological and experimental work. Special thanks go to Maja Felicia Bendtsen (Østkraft A/S) and her team for their hard work on the field, Dieter Gantenbein and his colleagues at IBM and Martin Bo Sjøberg and Andreas Arendt at Siemens for providing equipment and expertise, and finally Niels Ejnar Helstrup Jensen, Stig Holm Sørensen and Niels Per Lund from Energinet.dk for help with the data and constant feedback on this work.

REFERENCES

[1] "European Council Conclusions on 2030 Climate and Energy Policy Framework," European Council, Tech. Rep. October, 2014.

[2] N. O'Connell, P. Pinson, H. Madsen, and M. O'Malley, "Benefits and challenges of electric demand response: A critical review," *Renewable & Sustainable Energy Reviews*, vol. 39, pp. 686–699, 2014.

[3] Z. Xu, J. Ostergaard, and M. Togeby, "Demand as frequency controlled reserve," *IEEE Transactions on power systems*, vol. 26, no. 3, pp. 1062–1071, 2011.

[4] A. Yousefi, T. Nguyen, H. Zareipour, and O. Malik, "Congestion management using demand response and facts devices," *International Journal of Electrical Power & Energy Systems*, vol. 37, no. 1, pp. 78–85, 2012.

[5] J. Medina, N. Muller, and I. Roytelman, "Demand response and distribution grid operations: Opportunities and challenges," *IEEE Transactions on Smart Grid*, vol. 1, no. 2, pp. 193–198, 2010.

[6] R. Hledik and A. Faruqui, "Valuing demand response: International best practices, case studies, and applications," Brattle group website, 2015. [Online]. Available: http://www.brattle.com/system/publications/pdfs/000/005/343/original/Valuing_Demand_Response_-_International_Best_Practices__Case_Studies__and_Applications.pdf?1468964700

[7] G. Strbac, "Demand side management: Benefits and challenges," *Energy Policy*, vol. 36, no. 12, pp. 4419–4426, dec 2008.

[8] J. Torriti, M. G. Hassan, and M. Leach, "Demand response experience in Europe: Policies, programmes and implementation," *Energy*, vol. 35, no. 4, pp. 1575–1583, Apr. 2010.

[9] J. L. Mathieu, D. S. Callaway, and S. Kiliccote, "Examining uncertainty in demand response baseline models and variability in automated responses to dynamic pricing," in *IEEE Conference on Decision and Control and European Control Conference*. IEEE, Dec. 2011, pp. 4332–4339.

[10] O. Corradi, H. Ochsenfeld, H. Madsen, and P. Pinson, "Controlling Electricity Consumption by Forecasting its Response to Varying Prices," *IEEE Transactions on Power Systems*, vol. 28, no. 1, pp. 421–429, 2012.

[11] D. M. Keane and A. Goett, "Voluntary residential time-of-use rates: lessons learned from pacific gas and electric company's experiment," *IEEE transactions on power systems*, vol. 3, no. 4, pp. 1764–1768, 1988.

[12] P. Jazayeri, A. Schellenberg, W. Rosehart, J. Doudna, S. Widergren, D. Lawrence, J. Mickey, and S. Jones, "A survey of load control programs for price and system stability," *IEEE Transactions on Power Systems*, vol. 20, no. 3, pp. 1504–1509, 2005.

[13] V. Giordano, F. Gangale, G. Fulli, M. S. Jiménez, I. Onyeji, A. Colta, I. Papaioannou, A. Mengolini, C. Alecu, T. Ojala *et al.*, "Smart grid projects in europe: lessons learned and current developments," *JRC Reference Reports, Publications Office of the European Union*, 2011.

[14] P. Siano, "Demand response and smart gridsa survey," *Renewable and Sustainable Energy Reviews*, vol. 30, pp. 461–478, 2014.

[15] V. M. Balijepalli, V. Pradhan, S. Khaparde, and R. Shereef, "Review of demand response under smart grid paradigm," in *Innovative Smart Grid Technologies-India (ISGT India), 2011 IEEE PES*. IEEE, 2011, pp. 236–243.

[16] S. B. Harvey, "The Harvard Medical School Guide to Men's Health," *Publishers Weekly*, vol. 249, no. 31, 2002.

[17] P. Bartholomew, W. Callender, C. Hindes, C. Grimm, K. Johnson, M. Straub, D. Williams, M. Williamson, D. Hayes, W. Johnson, B. Nix, J. Lynch, and S. Romer, "Demand response measurement & verification," AEIC website, 2009. [Online]. Available: http://aeic.org/wp-content/uploads/2013/07/AEIC-MV-Whitepaper-Rev-051613.pdf

[18] J. Kwac, J. Flora, and R. Rajagopal, "Household energy consumption segmentation using hourly data," *Smart Grid, IEEE Transactions on*, vol. 5, no. 1, pp. 420–430, 2014.

[19] S. Haben, C. Singleton, and P. Grindrod, "Analysis and clustering of residential customers energy behavioral demand using smart meter data," *Smart Grid, IEEE Transactions on*, vol. in press, 2015.

[20] M. Chaouch, "Clustering-based improvement of nonparametric functional time series forecasting: Application to intra-day household-level load curves," *Smart Grid, IEEE Transactions on*, vol. 5, no. 1, pp. 411–419, 2014.

[21] F. Quilumba, W.-J. Lee, H. Huang, D. Wang, and R. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *Smart Grid, IEEE Transactions on*, vol. 6, no. 2, pp. 911–918, March 2015.

[22] "EcoGrid EU Website." [Online]. Available: www.eu-ecogrid.net

[23] Y. Ding, S. Pineda, P. Nyeng, J. Østergaard, E. M. Larsen, and Q. Wu, "Real-Time Market Concept Architecture for EcoGrid EU - A Prototype for European Smart Grids," *IEEE Transactions on Smart Grid*, vol. 4, no. 4, pp. 2006–2016, 2013.

[24] N. E. Helstrup, P. Lund, M. F. Bendtsen, D. Gantenbein, and A. Arendt, "Deliverable 6.3 - System operation and monitoring: Large-scale smart grids demonstration of real time market-based integration of DER and DR," EcoGrid EU Internal Report, Tech. Rep., 2013.

[25] G. Chicco, R. Napoli, and F. Piglione, "Comparisons among clustering techniques for electricity customer classification," *IEEE Transactions on Power Systems*, vol. 21, pp. 933–940, 2006.

[26] R. Xu and D. C. Wunsch II, *Clustering*. Wiley-IEEE Press, 2008, no. August.

[27] L. Kaufman and P. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," *Journal of the American Statistical Association*, vol. 86, no. 415, pp. 830–832, 1991.

[28] M. Zepeda-Mendoza and O. Resendis-Antonio, "Hierarchical Agglomerative Clustering," in *Encyclopedia of Systems Biology SE - 1371*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. Singer New York, 2013, pp. 886–887.

[29] B. Efron, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, vol. 1, no. 1, pp. 54–75, 1986.

[30] R. Yaffee, "Introduction to time series analysis and forecasting with applications of SAS and SPSS," *International Journal of Forecasting*, vol. 17, no. 2, pp. 301–302, 2001.

[31] P. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences*, pp. 49–55, 1936.

[32] Z. Prekopcsák and D. Lemire, "Time series classification by class-specific Mahalanobis distance measures," *Advances in Data Analysis and Classification*, vol. 6, no. 3, pp. 185–200, jul 2012.

[33] O. Ledoit and M. Wolf, "Honey, I Shrunk the Sample Covariance Matrix," pp. 110–119, 2004.

[34] J. Schaefer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications In Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

[35] J. P. Benzécri, *Analyse des données, Tome 1, La Taxinomie*, 1st ed. Dunod, 1976, [in french].

[36] C. Romesburg, *Cluster analysis for researchers*. Lulu. com, 2004.

[37] C. Meinert and C. Buck, *Clincal-Trials Design, Conduct, and Analysis*. Oxford University Press, USA, 1987, vol. 78, no. 4.

[38] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, 1st ed. CRC press, 1993.

[39] R. R. Wilcox, *Applying contemporary statistical techniques*. Gulf Professional Publishing, 2003.