# Cost-of-service segmentation of energy consumers

Adrian Albert and Ram Rajagopal

***Abstract*—Uncertainty in consumption is a key challenge at energy utility companies, which are faced with balancing highly stochastic demand with increasingly volatile supply characterized by significant penetration rates of intermittent renewable sources. This paper proposes a methodology to quantify uncertainty in consumption that highlights the dependence of the cost-of-service with volatility in demand. We use a large and rich dataset of consumption time series to provide evidence that there is a substantial degree of high-level structure in the statistics of consumption across users which may be partially explained by certain characteristics of the users. To uncover this structure, we propose a new technique for extracting typical *statistical signatures* of consumption - *energy demand distributions* (EDDs) - that is based on clustering distributions using a fast, approximated algorithm. We next studied the factors influencing the choice of consumption signature and identify certain types of appliances and behaviors related to appliance operation that are most predictive. Finally, we comment on how structure in consumption statistics may be used to target residential energy efficiency programs to achieve greatest impact in curtailing cost of service.**

***Index Terms*—Smart meter data, segmentation, service cost.**

## I. Introduction

With the increasing penetration of renewables on the grid, energy generation is becoming more volatile, and utility companies are turning to demand-side management as a cost-effective way to provide flexibility [1]. There is strong interest at many utilities to use the granular consumption data being collected by *smart meters* to develop tailored Energy Efficiency (EE) programs for the different segments of their customer base, such as targeting rebates for efficient appliances to the *right* users. But who are these users, and what do they respond to?

Operational decisions in the energy market depend heavily on the ability of the system operator to understand the risks (uncertainty) structure in supply and demand. In this work we propose a methodology to characterize demand uncertainty that is based on statistical signatures of consumption which we call *energy demand distributions (EDDs)*. EDDs offer a compact way to describe and compare the consumption of large populations of entities (here residential households, henceforth referred to using the generic term *user*) for the purpose of customer segmentation and marketing program targeting, while retaining a strong relationship to the *cost-of-service* on the grid. Moreover, the EDD is directly related to the load duration curve of a typical load. Our main hypotheses are that *i)* the consumption of a user population may be reasonably described with a small number of typical statistical signatures and *ii)* these patterns may be explained in part by certain user attributes.

Differences in aggregate consumption at the user level arise because of appliances, house characteristics, weather patterns, demographics, etc. [1]. The typical approach to modeling energy demand has previously been to implicitly assume normally-distributed aggregate loads (e.g., [2], [3]), and explain the average consumption and its variance. Our work takes a first step to explain the structure of the full distribution of energy demand - the EDD signature - as a function of user characteristics, with the purpose of differentiating users for energy efficiency interventions.

A first contribution of this paper is a framework to mine patterns in distributions of energy demand using a statistically-meaningful distance metric $d_{KS}$. We use simple machine learning techniques to derive a fast, approximate `K-Medians` algorithm [4] that may be used to obtain EDD clustering solutions on large data sets. To illustrate our technique, we show that the consumption of $\sim$ 900 real users may be described by with good confidence by 13 classes.

Our second contribution is to explain the choice of EDD signature in terms of characteristics of the users using discrete choice analysis [5] on data on more than 100 survey questions on demographics, appliance stock, attitudes toward energy efficiency etc. As such, we may use EDD class membership as "detector" to infer which users are more likely e.g., to be using large appliances such as clothes dryers, since those users may be good targets for energy efficiency rebates.

The rest of the paper is structured as follows. In Section II we describe the problem set-up. Section III reviews relevant prior literature. In Section IV we describe an algorithm we developed to mine EDD signatures from data. In Section V we describe the data used and the obtained typical EDD signatures. In Section VI we identify user features that determine typical EDD signatures. We conclude in Section VII.

## II. Patterns in Power Demand Distributions

### A. Problem statement

**Energy Demand Distributions.** We observe $N$ electricity consumption time series $L^i(t)$ for $t = 0, ..., T$, each representing the demand of electricity of an individual user $i$ ($i = 1, 2, \ldots, N$) over the period $T$. From the standardized data, we build the empirical cumulative distribution function (CDF) of consumption $F_L^i(l) = \frac{1}{T}\sum_{t=0}^{T} \mathbf{1}(L^i(t) \leq l)$ ($l$ is

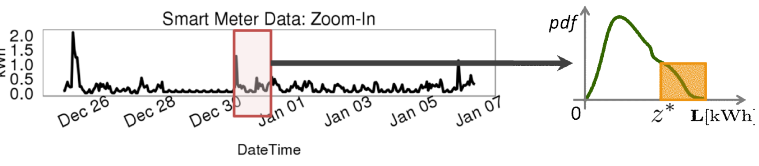


Fig. 1. Computing energy demand distributions for a unit of analysis (shaded in red, e.g. overall, time-of-day). A time series (left) is then represented as a statistical signature (right).

A. A. is with Electrical Engineering at Stanford.
R. R. is with Civil and Environmental Engineering at Stanford.

an arbitrary consumption level) and the associated probability distribution function (pdf) $f^i(l)$, as shown in Figure 1. Note that the load duration curve for the random variable $\mathbf{L}$ is calculated as $\bar{F}^i_L(l) = 1 - F^i_L(l)$. This operation may be performed at the level of analysis desired, e.g., the entire time series (as done here) or by hour-of-day. Typical consumption pdfs - here called *energy demand distribution (EDD)* - may serve as basic statistical signatures summarizing consumption of a given user.

**EDD classes.** Given $N$ pdfs we are interested in identifying a small number $K$ ($K \ll N$) of groups such that the distributions in any given group are statistically similar to each other, and statistically different from the ones in other groups. These clusters represent a segmentation of the user base by their cost-of-service. Clustering requires pairwise comparisons of EDDs in a statistically-meaningful way; for this we have developed the $d_{KS}$ distance (see Section II-D).

**Determinants of EDD class.** We study whether there are certain user characteristics (behaviors, appliances etc.) that affect the choice of EDD patterns - i.e., all things equal, what might a system operator attempt to change in order to improve consumption statistics of a certain type of users? We address this using a discrete-choice multinomial logit formulation [5].

### B. The cost of servicing a consumer

The current structure of the energy market requires that utilities procure energy in advance to satisfy demand in aggregate. Electricity is cheaper for less variable aggregate demands. But, the cost of variability in consumption is borne equally by all customers, although a few may have much larger variability. Identifying those consumers is important since they are ones who can benefit most from targeted interventions.

Consider the simplified model of an electricity market in Figure 2 as in [6]. The market has two stages: day ahead and real time. The consumer uses a random amount of energy $\mathbf{L}$ [MWh] revealed only in the real-time. The utility procures energy $z$ [MWh] in the day ahead at a cost $p$ [$/MWh], and cost $q$ [$/MWh] in real-time. Note that $q > p$, since otherwise the utility would always buy energy real-time as opposed to planning ahead (the average values in California are $p = \$52$/MWh and $q = \$70$/MWh [6]). It can be shown (see Appendix A) that the optimum purchase $z^*$ is

$$F_L(z^*) = \frac{q-p}{q}, \tag{1}$$

which depends directly on the CDF $F_L$. The expected cost to the utility to serve this consumer is

$$C^* = p\ \mathbb{E}[\mathbf{L}|\mathbf{L} \geq z^*] \tag{2}$$

(in [$/MWh]), i.e., the cost-of-service is driven by the tail end of the distribution shaded in the right panel in Figure 1. As such, different *shapes* in distributions carry different costs of service, even as users may look identical when described by traditional measures employed in the industry (means and variances). Conversely, relative changes in mean and variance of demand (i.e., shifting and scaling operations on $\mathbf{L}$) for fixed market conditions $(p, q)$ only select levels of $z^*$ from the same underlying distribution shape $f$. In an operational setting, equation (2) also implies that demand forecasting methodologies that focus on point-wise prediction of future mean and spread of consumption may fail to address their very purpose of reducing financial uncertainty, since the structure of the variability $f$ is a key driver of cost.

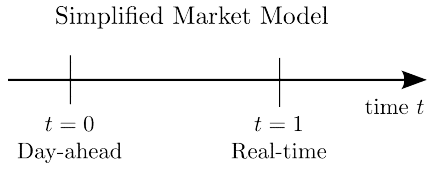


Fig. 2. Simplified two-stage model of an energy market.

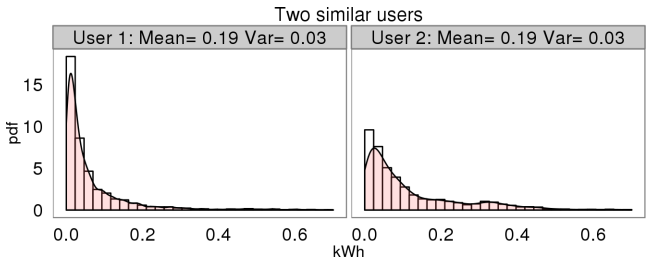


Fig. 3. Consumption distributions (pdfs) of two real users having the same means and variances, but different mass in the tails. Estimated costs-of-service $C^*$ are $7.5/MWh (left user) and $10.8/MWh (right user).

### C. EDD classes

Typical strategies in modeling consumption decisions build regression models for which the mean of consumption is modeled a function of exogenous parameters (e.g., weather [3]), and the error term follows a zero-mean, constant variance Gaussian. However virtually no previous studies have focused on the structure of the full distribution of the random component, which is the main driver of risk - and thus cost-of-service.

Consider the consumption distribution functions of two real users in Figure 3. The two users' consumption is identical when compared using the first two moments of the distribution (mean and variance), yet clearly it differs when comparing the full distributions. As argued above, the different distribution signatures yield different costs to the utility of servicing these two users (with User 2 being more expensive to service, as more of the mass in his pdf is in the tail).

Thus, our proposed model of consumption assumes that the observed usage - denoted by $\mathbf{Y}$ in Figure 4 (left) - is the result of three user choices: *i)* the average consumption (mean) $\mu$, *ii)* a consumption variance $\sigma^2$, and *iii)* a distribution prototype given by the EDD signature $f$. Given two loads described by the random variables $\mathbf{X}$ (with mean $\mu_X$ and variance $\sigma^2_X$) and $\mathbf{Y}$ (with mean $\mu_Y$ and variance $\sigma^2_Y$), we consider them similar (i.e., they belong to the same EDD *class*) if

$$\mathbf{Y} = a\mathbf{X} + b, \tag{3}$$

where we call $a$ the *scale parameter* and $b$ the *shift parameter* (see left panel in Figure 4). This affine transformation allows to directly relate and recover the statistics of $\mathbf{Y}$ from those of $\mathbf{X}$ in a straightforward manner, as $F_{Y;a,b}(y) = F_X\left(\frac{y-b}{a}\right)$. Note that (3) defines similarity between statistics of $\mathbf{X}$ and $\mathbf{Y}$ that is agnostic of the *relative* mean and variance of the two random

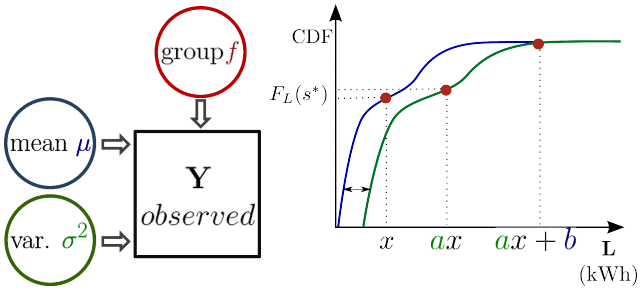


Fig. 4. *Left:* user consumption model consisting of choice of mean $b$, variance $a$, and signature group $f$. *Right:* illustration of the shift and scale operations.

variables (with $\mathbb{E}[Y] = a\,\mathbb{E}[X] + b$ and $var(Y) = a^2 var(X)$). Were $\mathbf{X}$ and $\mathbf{Y}$ Gaussian, the relative mean and variance are enough to characterize the degree of similarity between their distributions, which is captured as a special case in (3).

### D. The $d_{KS}$ distributional metric

For clustering a direct way of comparing two distributions $F_X$ and $F_Y$ is needed that carries statistical meaning. Here we use the two-sample *Kolmogorov-Smirnoff (K-S) statistic*

$$KS(F_X, F_Y) = \sup_x |F_X(x) - F_Y(x)|. \tag{4}$$

This is a non-parametric alternative to the more widely-used tests (such as the $t$-test) that make parametric assumptions on the distributions involved. We chose this test because of *i)* distribution functions in our specific application (energy demand distributions) are very likely to violate the normality assumption involved in most statistical tests, as indicated by empirical analysis on real data presented in this paper; and *ii)* it is easy to compute and has useful properties for our modeling needs (see brief discussion in Appendix C and a detailed exposure in e.g., in [7]). We wish to assess the distributional distance between a given distribution $F_X$ and a distribution $F_{Y;a,b}$ to which a transformation as in Equation (3) has been applied. Moreover, in comparing the two distributions under our model (3) we seek the optimum transformation (i.e., the optimum scale parameter $a^*$ and shift parameter $b^*$) such that:

$$d^*_{KS}(F_X, F_Y) = \min_{a,b} \sup_x \left| F_X(x) - F_Y\left(\frac{x-b}{a}\right)\right| \tag{5}$$

In effect, we require extracting (or adding) the optimum relative proportion from the mean and variance of $\mathbf{X}$ such that the maximum amount of probability mass is matched with the mass concentration of $\mathbf{Y}$. We show in the Appendix C that $d_{KS}$ is a metric and therefore desirable for a clustering application, and discuss how we computed it numerically.

### E. Discussion of model assumptions and limitations

For the sake of simplicity our model of consumption relies two assumptions that we discuss here. First, utility companies generally bid on forecasts of *aggregate* demand, not individual users. For the EDD shapes to encode the cost-of-service of the group aggregate demand, all users would have to consume exactly the same at every given time - i.e., they would have to be perfectly correlated. However the framework proposed here is useful for *comparing* the relative impact on the grid across users whose consumptions differ in variability.

Second, describing users' consumption through their EDD alone assumes time-independent, i.i.d. demand that does not capture serial correlations. This would be a notable limitation if our purpose was to build detailed models for forecasting individual consumption; however our purpose is to extract useful statistical benchmarks from arbitrary time series. These time series may be constructed for each hour of the day (or day of the week etc.). That is, as a first step into that direction, one may divide up the original time series $\{L_t\}$ into hourly blocks, compute CDFs of consumption $F^h_L(l)$ $(h = 1, ..., 24)$, and identify typical statistical signatures for each hour of the day. Moreover, EDD shapes may be computed by pooling or aggregating data across users, which reveals variability structures that are persistent over time, as well as reduce the number of samples size necessary for estimating the distributions. In addition, our method significantly generalizes the process of determining typical load duration curves for individual customers. Such an approach will be required for designing novel tailored pricing contract mechanisms for classes of consumers.

## III. Related Literature

Most traditional segmentation studies in the energy sector have previously focused on "psychographic" analysis of customer attitudes toward energy consumption (e.g., [8]), as opposed to the users' actual consumption data. Data-driven segmentation studies have focused on forecasting and clustering of daily load profiles for small to medium-sized time series datasets, most often for aggregate loads at different levels (substations or buses). The typical application (e.g., [9], [10], [11], [12], [13]) is to group raw time series or features extracted from this data using standard algorithms (i.e., `K-Means`) and distance metrics (i.e., Euclidean). A review of clustering techniques for daily consumption profiles is given in [14]. In contrast, to the best of our knowledge this paper is the first to explore in depth patterns in statistical signatures of energy demand, greatly expanding on our preliminary previous work [15]. Moreover, as opposed to previous studies that investigated statistics in demand (such as [16]) using only a few users, we extract non-parametric signatures from a large user sample.

The idea that non-Gaussian distribution signatures carry information due to their shape types has been studied in other contexts as well. In [17] a goal similar to ours is pursued - clustering financial loss distributions into a small number of prototypes - although with a different approach that requires *all* data samples to be available to the user (as opposed to just a representation of the CDF as in our case) and with a different statistical metric (the Anderson-Darling test). A related recent study in this context is [18], where the authors cluster (in a hierarchical fashion) operational loss data into homogeneous classes using the Kolmogorov-Smirnov statistic to capture similarity between distributions. In contrast to their approach, our affine model (3) allows for shifting and scaling to further increase data compression, and our accelerated algorithms allow clustering of much larger datasets.

## IV. Mining EDD signatures from data

### A. Fast, approximate K-Medians clustering

We extract EDD signatures using an extension of the `K-Medians` algorithm [4]. The main advantages of this approach over the popular `K-Means` algorithm are *i)* we use the statistically-meaningful distribution distance (5) instead of the Euclidean distance, and *ii)* we avoid using an averaging operation for computing a cluster centroid: as the CDFs are already an integrated, smooth signal, and averaging yields smoothed-out, uninformative cluster centers.

We seek to separate the initial set $\mathcal{A}$ of users (represented by their CDFs $F_i$, $i = 1, ..., N$) into $K$ (disjoint) clusters $\mathcal{C}_k$, $k = 1, ..., K$, such that $\mathcal{A} = \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_K$. Each cluster $k$ may then be represented by a prototype $\hat{F}_k \in \mathcal{C}_k$. Cluster membership is encoded using an assignment function $C: \ F_i \in \mathcal{C}_k \rightarrow C(i) = k$. The algorithm starts with an initial random assignment $C_0$ and iteratively minimizes the overall inhomogeneity objective

$$\min_{\mathcal{C}} \sum_{k=1}^{K} \sum_{F_i \in \mathcal{C}_k} d_{KS}(F_i, \hat{F}_k) \tag{6}$$

in a two-step process as described in Algorithm 1.

**Algorithm 1** Accelerated `K-Medians` algorithm.

**Input:** $K$; CDFs set $\mathcal{A} = \{F_i\}_{i=1,...,N}$; initial assignment $C_0$.
**Output:** Cluster assignment $C$ and centers $\{\hat{F}_k\}_{k=1,...,K}$.

Repeat until $C$ does not change:

1) **Classification step.** Find the median of each cluster $k$ (the element $\hat{F}_k$ for which the total distance from all the other CDFs in the cluster to $\hat{F}_k$ is minimized), with $\mathcal{C}_k = \{F_i \in \mathcal{A} | C(i) = k\}$:

$$\hat{F}_k = \underset{\{F_{i'} \in \mathcal{C}_k\}}{\operatorname{argmin}} \sum_{F_{i'} \in \mathcal{C}_k} d_{KS}(F_i, F_{i'}) \ \ \forall k = 1, ..., K. \tag{7}$$

2) **Assignment step.** For each $F_i \in \mathcal{A}$, find the cluster assignment $C(i)$ that minimizes total K-S distance from each CDF to the closest cluster center:

$$C(i) = \underset{1 \leq k \leq K}{\operatorname{argmin}} \ d_{KS}(F_i, \hat{F}_k) \tag{8}$$

This operation also results in a set of optimal shift and scale parameters $a_i^*$ and $b_i^*$ from point $F_i$ to the respective cluster center indicated by $C(i)$.

A brute-force algorithm will generally have a quadratic time complexity in the size $N$ of the data set (more precisely, $\mathcal{O}\left(i\frac{N^2}{K}\right)$, where $K$ is the number of clusters and $i$ is the number of iterations until convergence), as the median computation in the *Classification step* requires calculating all pairwise distances. We use results reported in the machine learning literature to accelerate computation in both steps.

**Assignment step.** To accelerate this step, we employ a simple technique introduced in [19] to decrease the number of comparisons necessary to arrive at a provable exact assignment solution (as compared with the naive algorithm) by using the metricity (symmetry and triangle inequality) of $d_{KS}$.

**Classification step.** Note that for `K-Means` the *Classification step* is straightforward to compute, as it only involves an averaging operation. However for `K-Medians` this is not true anymore, as the new centroid is the median, whose search requires computing all pairwise distances between points in the cluster. To accelerate the *Classification step* we use a straightforward, yet effective randomized technique proposed in [20] that computes an approximate median of a set with significant reduction in the number of necessary distance computations. This approach is based on sampling the data set, computing partial sums (6) only for points in the sample, then computing a median solution from eligible full sums in the selected sample. By selecting a reasonable value for the sample size (here $\sim \log N$ from numerical experiments), we obtain sizable performance gains at low median objective error.

**Selecting model size $K$.** Up to now we have assumed $K$ to be known; however this is not the case in real applications. To identify an acceptable value for $K$ that trades off parsimony and low error, we start from the observation [21], [4] that clustering quality (as measured by the objective (6)) does not change dramatically when $K$ is increased past its optimum value. A typical profile of (6) is depicted in Figure 6, which was obtained by computing clustering solutions for the $N \sim 900$ overall EDDs in our dataset and $K$ between $K_{\min} = 2$ and $K_{\max} = 2^7$. We used a binary search procedure as outlined below [21]:

1) Compute clustering solutions for $K = 2, 2^2, 2^3, ..., 2^v, ...$ until the median cluster radius $f(K) \geq c_\alpha$ (with $c_\alpha = 0.075$ a value chosen by the user) at step $v$; then our desired value of $K$ must lie between $\eta = 2^{v-1}$ and $\xi = 2^v$;
2) Set $r = \frac{\eta+\xi}{2}$ and compute $f(r)$ by running Algorithm 1;
3) If $f(r) > c_\alpha$, set $\xi = r$, else set $\eta = r$;
4) Repeat steps *2* and *3* until $\eta = \xi$.

This binary search procedure (steps marked in Figure 6) avoids computing clustering solutions for most $K$ values in the search interval, and will identify a reasonable model size $K^*$ in $\log_2(K_{\max} - K_{\min})$ steps ($5-8$ steps in our expriments).

### B. Algorithm performance

The time complexity of the *accelerated* `K-Medians` is $\mathcal{O}\left(N \log \frac{N}{K}\right)$, which makes our technique suitable for large datasets. The performance is documented in Figure 6. In the top panel, we compute clustering solutions for samples of increasing size $N$ drawn at random from the overall demand distributions in the Google dataset [3] (we keep the model size constant, $K = 10$). The number of distance evaluations is plotted against $N$ for our EDD clustering technique (blue line), a reference curve for naive `K-Medians` of order $O(\frac{N^2}{K})$, and a reference curve for $O(N \log \frac{N}{K})$. Moreover, in our experiments the solution quality (as measured by the clustering objective 6) is barely affected, with relative errors (compared with the naive algorithm) fluctuating around $3 - 4\%$ (because of locality of solutions given by `K-Means`).

## V. Mined EDD signature classes

### A. Data description and setup

We used two types of data collected from Google employees through an 8-month (Mar.-Oct. 2010) experiment [3]:

1) **Energy demand time series** of 10-minute resolution for about 1100 users.

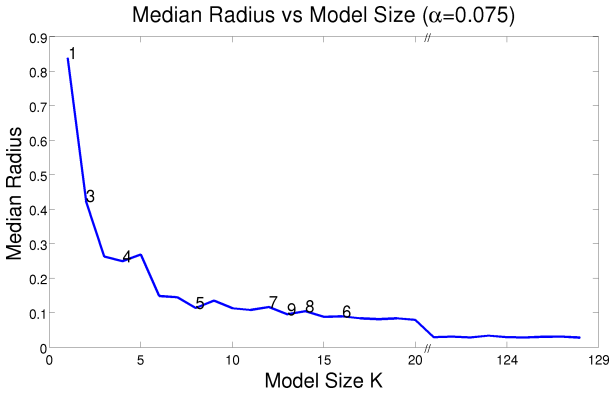


Fig. 5. Computing the optimum model size $K^*$. Steps in the binary search are marked on the objective profile.

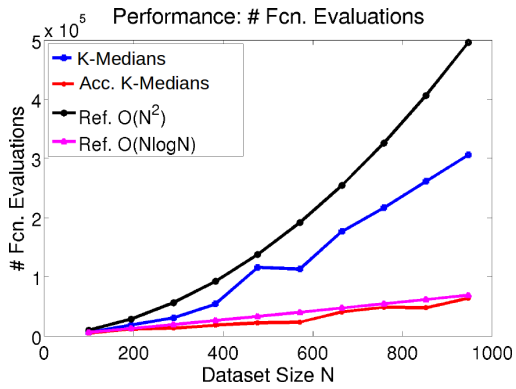


Fig. 6. Performance of accelerated `K-Medians` (blue) is empirically similar to $O(N \log N)$ (magenta).

2) **Socio-economic data** obtained via an on-line survey from about 950 participants.

With penetration rates of over 70% in states like CA, consumption data as we use here is becoming ever more ubiquitous at utilities. Typical smart meters record data every 15 minutes or hour, although that is largely an operational decision from the part of utilities (because of data transmission, storage and privacy concerns, etc.) rather than a hardware constraint [22].

The survey data we employ is rather detailed - over 100 variables on demographics, lifestyle, attitudes and behaviors towards energy efficiency - but utilities are starting to collect such information at fast pace, e.g., from the U.S. Census or the American Community Survey. A detailed discussion of this data is not possible here, but is given in [3], [2]. Most users are engineers located on the U.S. West Coast, are generally well-off, and live in medium and large houses or apartments.

The sensor data was unfortunately plagued by reading errors and required extensive cleaning. We only retained the time series that had at least a week's worth of measurements ($\geq$ 1500 samples). The final dataset had $N = 898$ users.

### *B. Mined EDD classes and cost-of-service segmentation*

The EDD clustering solution obtained on the data described above is presented in Figure 7. The top panel illustrates obtained cluster medians; the bottom panel presents a top-view on all the distributions in the respective clusters (each row in the heat map color-codes a distribution, such that blue is low pdf value, and red is high pdf value). The heatmap offers a concise visualization of clustering quality, as it is clear that each cluster contains pdfs that are well aligned in shape. For a median cluster radius of $\alpha = 0.075$, the $\sim$ 900 overall distributions were clustered in 13 groups, which yields an average within-cluster dispersion of $\sim 0.05$, as indicated in Figure 8.

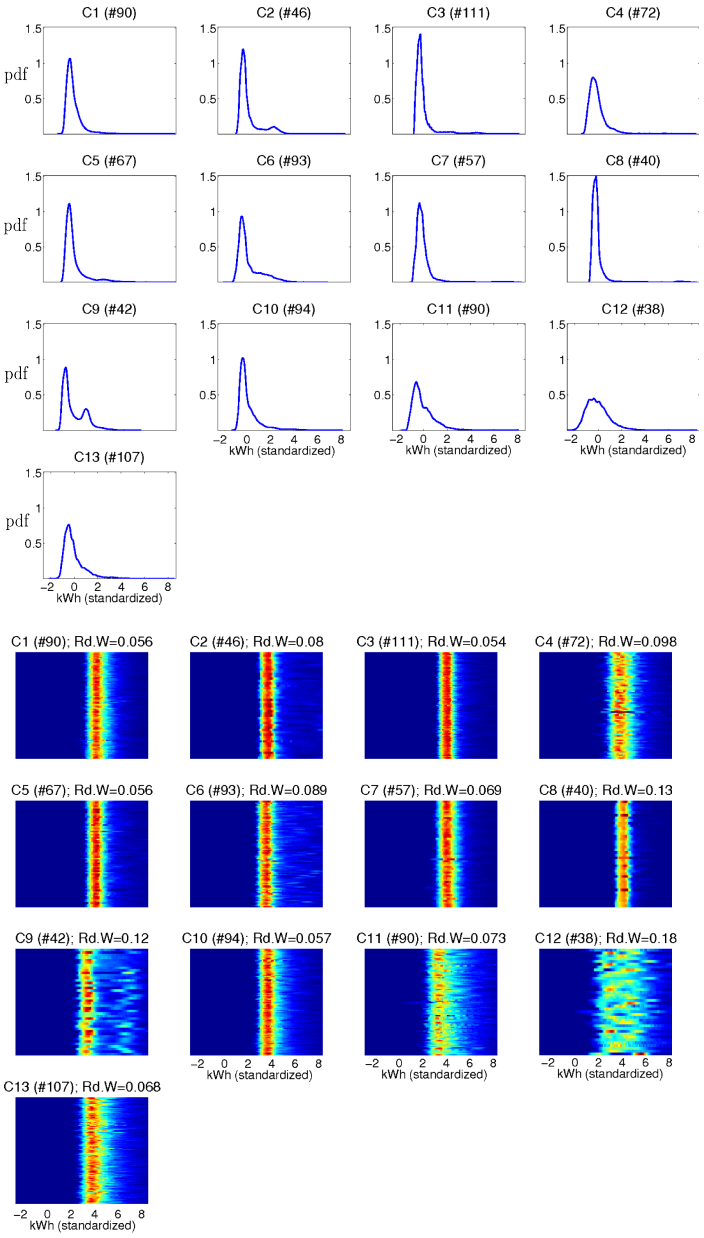


Fig. 7. Typical signatures mined from $\sim$ 900 overall EDDs. *Top:* Cluster medians (as pdfs); *Bottom:* Cluster membership as heatmaps (each row color-codes a pdf in a cluster, with a shift and scale transformation (3) applied to all cluster members to their cluster center). The tight visual structure suggests that clustering indeed uncovers patterns that are similar within each EDD class. In either figure, all sub-panels share the vertical and horizontal axes.

Based on Equation (2) we compute the cost $C^*$of servicing users that consume according to the same mean and variance, but with different EDD signatures. The cost of a EDD signature then represents the scaling factor that may be used together with a given mean and variance to compute cost of service for a particular user segment. For the values of $p$ and $q$ in Section

| Variable | C.2 | C.3 | C.4 | C.5 | C.6 | C.7 | C.8 | C.9 | C.10 | C.11 | C.12 | C.13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Has Spa or Pool Heater? | -0.00 | 0.02 | 0.01 | 0.03 | 0.10 | 0.02 | -0.00 | 0.25 | -0.02 | 0.06 | 0.30 | 0.06 |
| No. AC units | -1.22 | 1.37 | 1.67 | 1.91 | 1.60 | 1.88 | -1.05 | -0.78 | 1.16 | 3.81 | 1.88 | 2.82 |
| Has clothes dryer? | -0.89 | 0.69 | 0.58 | 0.76 | 0.37 | 0.50 | -0.85 | -0.40 | 0.10 | 0.76 | 0.34 | 0.64 |
| Has programm. thermostat? | 0.10 | 0.08 | 0.10 | 0.03 | 0.16 | -0.09 | 0.07 | 0.36 | -0.02 | 0.06 | -0.05 | 0.01 |
| Has electric water heating? | -0.37 | 0.16 | 0.08 | 0.13 | 0.08 | 0.05 | -0.19 | -0.11 | 0.05 | 0.06 | 0.04 | 0.05 |
| Occupants over 65? | -0.00 | 0.02 | 0.02 | 0.03 | 0.00 | 0.04 | -0.00 | -0.00 | -0.03 | 0.03 | 0.00 | 0.03 |
| No. occupants | -1.34 | 1.32 | 1.34 | 1.36 | 0.45 | 1.11 | -1.22 | -1.26 | 0.26 | 1.37 | 0.09 | 1.36 |
| Floor area | -5.04 | 4.35 | 4.54 | 4.51 | 2.36 | 3.89 | -4.68 | -0.75 | 0.78 | 5.24 | 1.40 | 4.70 |
| Is hot water part of rent? | -0.06 | 0.13 | 0.12 | 0.22 | 0.07 | 0.05 | -0.19 | 0.00 | 0.04 | 0.11 | 0.06 | 0.09 |
| Turns off computers? | -0.07 | 0.09 | 0.10 | -0.03 | -0.05 | 0.16 | -0.17 | 0.19 | 0.10 | -0.04 | -0.04 | 0.06 |
| Uses AC efficiently? | -0.53 | -0.28 | -0.41 | -0.10 | -0.01 | -0.51 | -0.03 | -0.16 | 0.01 | -0.20 | -0.14 | -0.18 |

TABLE I
MNL ELASTICITIES (REFERENCE CLASS 1) FOR SEVERAL REPRESENTATIVE CHARACTERISTICS.

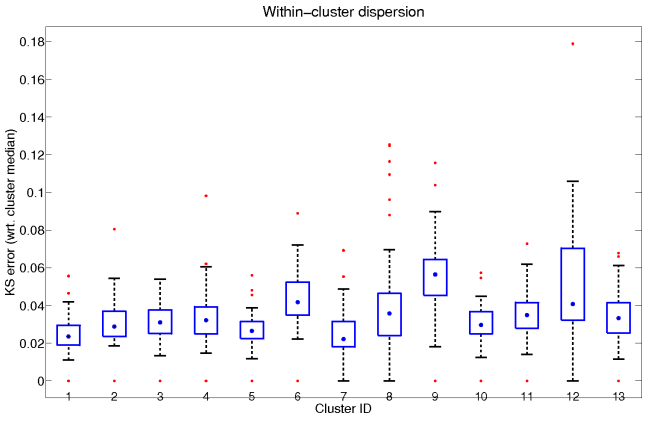


Fig. 8. Average cluster dispersion for $K = 13$.

| C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 7.5 | 8.2 | 7.3 | 10.1 | 8.3 | 10.3 | 6.9 | 5.8 | 14.2 | 9.9 | 13.8 | 16.6 | 12.3 |

TABLE II
COST-OF-SERVICE FOR EDD SIGNATURES ($/MWH).

II, typical cost-of-service for different EDD classes are given in Table II. A further classification yields:

*"Energy heroes":* low $C^*$ ($< 7.5$$/MWh) in clusters C1, C3, C7, and C8. These users are already economical to service, so they are not effective program targets (if targeting is costly).

*"Energy delinquents":* high $C^*$ ($> 12$$/MWh) in "fat-tail" clusters C9, C11, C12, and C13. These users have a highly variable consumption; as it might be too hard or costly to affect their consumption through efficiency programs.

*"Potential targets":* medium $C^*$ ($8 < C^* \leq 12$ $/MWh) in clusters C2, C4, C5, C6, C10. These users are the "low hanging fruit" for targeting programs: changing their usage might be not too difficult, and it would have a sizeable impact on the system, as they comprise $\sim 45\%$ of the sample.

We note that beside cost-of-service, another relevant interpretation of EDD classes is that they allow describing different load durations for users with the same means and variances. As such, multiple-peak shapes may represent activities that could be shifted away through appropriate interventions (e.g., pricing mechanisms). Within each class, users with higher means and variances can serve as potential targets for load reduction.

## VI. DETERMINANTS OF EDD CLASS

We argue that some characteristics of users carry greater weight than others in predicting whether users consume in similar ways. In essence, if EDD signatures encode information about user characteristics, they may serve as proxy for program targeting. Since prediction at the individual level offers little insight into more general trends over a population, we characterize the *expected shares* of the population consuming according to each EDD class.

**Finding:** *The largest determinants of change in EDD class shares are the presence of large fixtures (such as AC and clothes dryer) and user and occupancy characteristics (floor area, number of inhabitants). Behavior change results in relatively smaller change in shares of different classes.*

For this we used a multinomial logit *MNL* model as described in Appendix D. Table I presents MNL group-average elasticities for the most important variables. Note that variables such as `Floor area`, `No. AC units`, and `No occupants` all exhibit strong effects. However variables affect class shares in different ways: an increase in the floor area will decrease the shares of EDD classes 2 and 8 (strong negative elasticities), while increasing the shares of classes 11 and 13. The results in Table I may be used as a starting point to guide targeting programs that seek to change the proportion of people who consume according to "undesired" patterns. For example, one may wish to reduce the occurrence of the double-peaked EDD 9; among the interventions that one might try, a program that makes people more likely to install programmable thermostats will have an effect twice as strong than one that educates them about turning off appliances. If one wants to increase the share of the comparatively "well-behaved" EDD class 7, one way would be to determine people to reduce the number of clothes dryers or AC systems, noting however that the latter intervention is about $50\%$ more effective than the former.

## VII. CONCLUSIONS AND FUTURE WORK

This paper developed a methodology - *Energy Demand Distributions* (EDD) - for clustering univariate probability distributions that encode the characteristic cost-of-service of individual users. We applied the framework to segmenting a large

population of users by their variability signatures. We introduced a statistically-meaningful distance metric $d_{KS}$ between two distributions, and used it to develop a fast, approximated clustering algorithm. Moreover, we analyzed the factors that determine the choice of EDD signatures using a comprehensive survey, and show that the presence and operation of certain large appliances has a stronger effect in driving variability than self-reported attitudes on energy use.

The present work did not address time and geography-based correlations in demand - both key determinants of peak load buildup - which we are addressing in a forthcoming study.

## ACKNOWLEDGMENTS

We thank James Sweeney, Pravin Varaya, Duncan Callaway, Raffi Sevlian, and Sam Borgeson for comments on an early version of the manuscript; and Carrie Armel and ARPA-E for providing the dataset. A.A. is supported financially by the Precourt Energy Efficiency Center. R.R. is supported financially by the Powell Foundation, the TomKat Center, and ARPA-E.

## REFERENCES

[1] A. Faruqui, P. Fox-Penner, and R. Hledik. Smart grid strategy: Quantifying benefits. *Public Utilities Fortnightly*, July 2009.

[2] Amir Kavousian, Ram Rajagopal, and Martin Fischer. Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, 55(0):184 – 194, 2013.

[3] Anant Sudarshan June A. Flora Sebastien Houde, Annika Todd and K. Carrie Armel. Real-time feedback and electricity consumption: A field experiment assessing the potential for savings and persistence. *The Energy Journal*, 0(Number 1), 2013.

[4] T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, 2009.

[5] Moshe Ben-Akiva and Steven R. Lerman. *Discrete choice analysis*. MIT Press, 1985.

[6] E.Y. Bitar, R. Rajagopal, P.P. Khargonekar, K. Poolla, and P. Varaiya. Bringing wind energy to market. *IEEE Transactions on Power Systems*, 2011.

[7] Raul H C Lopes, Ivan Reid, and Peter R Hobson. The two-dimensional kolmogorov-smirnov test. *Monthly Notices of the Royal Astronomical Society*, 335(1):73–83, 2007.

[8] M. Pedersen. Segmenting residential customers: energy and conservation behaviors. Number 7, pages 229–241, 2008.

[9] M Espinoza, C Joye, R Belmans, and B DeMoor. Short-term load forecasting, profile identification, and customer segmentation: A methodology based on periodic time series. *IEEE Transactions on Power Systems*, 20(3):1622–1630, 2005.

[10] Christoph Flath, David Nicolay, Tobias Conte, Clemens Dinther, and Lilia Filipova-Neumann. Cluster analysis of smart metering data. *Business Information Systems Engineering*, 4(1):31–39, 2012.

[11] Teemu Räsänen and Mikko Kolehmainen. Feature-based clustering for electricity use time series data. In *Proceedings of the 9th international conference on Adaptive and natural computing algorithms*, ICANNGA'09, pages 401–412. Springer-Verlag, 2009.

[12] V. Figueiredo, F. Rodrigues, Z. Vale, and J.B. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *Power Systems, IEEE Transactions on*, 20(2), 2005.

[13] G.J. Tsekouras, N.D. Hatziargyriou, and E.N. Dialynas. Two-stage pattern recognition of load curves for classification of electricity customers. *Power Systems, IEEE Transactions on*, 22(3):1120–1128, 2007.

[14] G. Chicco, Roberto Napoli, and Federico Piglione. Comparisons among clustering techniques for electricity customer classification. *Power Systems, IEEE Transactions on*, 21(2):933–940, 2006.

[15] Adrian Albert, Ram Rajagopal, and Raffi Sevlian. Power demand distributions: Segmenting consumers using smart meter data. In *BuildSys/SenSys, Seattle, WA*, November 2011.

[16] E. Carpaneto and G. Chicco. Probabilistic characterisation of the aggregated residential load patterns. *Generation, Transmission Distribution, IET*, 2(3):373–382, 2008.

[17] Eric Cope. Modeling operational loss severity distributions from consortium data. *Journal of operational risk*, 5(4), 2011.

[18] Fabio Piacenza, Daniele Ruspantini, and Aldo Soprano. *Journal of Operational Risk*, 1(3):51–59, 2006.

[19] Charles Elkan. Using the triangle inequality to accelerate k-means. In Tom Fawcett and Nina Mishra, editors, *ICML*. AAAI Press, 2003.

[20] Luisa Micó and José Oncina. An approximate median search algorithm in non-metric spaces. *Pattern Recognition Letters*, 22(10), 2001.

[21] Anand Rajaraman and Jeffrey D Ullman. Mining of massive datasets. *Lecture Notes for Stanford CS345A Web Mining*, 2010.

[22] C. K. Carmel, G. Shrimali, and A. Albert. Disaggregation: the holy grail of energy efficiency? *Energy Policy*, 2012.

[23] Nicholas C. Petruzzi and Maqbool Dada. Pricing and the news vendor problem: a review with extensions. *Oper. Res.*, 47(2):183–194, 1999.

## APPENDIX

### A. Cost-of-service of a single consumer

The setup described above in Section II-B is a Newsvendor-type problem (see e.g., [23] for a review), where the vendor (the electric utility) operates in a two-stage market. At time $t = 0$, the utility purchases a quantity $z$ at price $p$ to service an unknown demand $l$ (described by the random variable $\mathbf{L}$ with known statistics $F_L(l)$) at time $t = 1$. If $z < l$ (real-time demand is higher than pre-ordered amount), the utility may purchase the difference on the spot market at price $q$ (with $q > p$). Thus the cost of servicing the consumer is:

$$C(z, \mathbf{L}) = q \max\{\mathbf{L} - z, 0\} + pz, \tag{9}$$

whose expected value may be written as (assuming $F_L(0) = 0$):

$$\begin{aligned}
\mathbb{E}[C(z)] &= \int_0^\infty q \left(\max\{x - z, 0\} + pz\right) dF_L(x) \\
&= q \int_z^\infty x dF_L(x) - qz \int_z^\infty dF_L(x) + pz \\
&= q \left(\bar{L} - zF_L(z) + \int_0^z F_L(x)dx\right) - \\
&\quad qz(1 - F_L(z) - F(0)) + pz \\
&= q(\bar{L} - z) + q \int_0^z F_L(x)dx + pz. \qquad (10)
\end{aligned}$$

In arriving at the third expression above starting from the second one we made use of an integration by parts, $\int_z^\infty x dF_L(x) = \bar{L} - \int_0^z x dF_L(x) = \bar{L} - xF(x)|_0^z + \int_0^z F(x)dx$. Then the optimum order quantity $z^*$ may be calculated from the first-order optimality condition:

$$\frac{\mathbb{E}[C(z)]}{dz} = -q + p + q \frac{d}{dz} \int_0^z F_L(x)dx = 0. \tag{11}$$

From the Leibniz rule of differentiation [23], the integral term above reduces to $F_L(z)$, and as we have for $z^*$:

$$F(z^*) = \frac{q - p}{q}, \tag{12}$$

and the associated cost-of-service follows as in Section II-B.

### B. The Kolmogorov-Smirnoff statistic $D_{mn}$

Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables having distribution functions $F_X$ and $F_Y$. Suppose $\{X_1, X_2, ...X_n\}$ and $\{Y_1, Y_2, ...Y_m\}$ are two samples (of respective sizes $n$ and $m$) coming from $\mathbf{X}$ and $\mathbf{Y}$, and that the empirical CDFs are $F_X^n$

and $F_Y^m$. Then we want to test

$$H_0 : F_X = F_Y \text{ vs. } H_1 : F_X \neq F_Y. \tag{13}$$

Define the *two-sample Kolmogorov-Smirnoff statistic* as

$$D_{mn} = \sqrt{\frac{mn}{m+n}} \sup_x |F_X^n(x) - F_Y^m(x)|. \tag{14}$$

It is shown in [7] that for continuous $F_X$ and $F_Y$ the distribution of (14) does not depend on either $F_X$ or $F_Y$.
Given a desired confidence $\alpha$, critical values $c_\alpha$ for which $H_0$ may be rejected under the KS test for different sample sizes (when $n = m$) are given in Table III. The decision rule is:

$$\delta = \begin{cases} H_0 & \text{if } D_{mn} \leq c_\alpha; \\ H_1 & \text{if } D_{mn} > c_\alpha. \end{cases} \tag{15}$$

Here we adopted the simplification $m = n = 1,500$, which is a typical number of weekly measurements (the lowest time disaggregation level that we considered). According to Table III, we thus require a value of $d_{KS}^* \equiv D_{nn} \approx 0.07$ to reject $H_0$ at a level of confidence of $0.001$. We adopt a very conservative stance to committing Type I errors (rejecting a true $H_0$).

| Sample Size n=m/ $\alpha$ | 0.1 | 0.05 | 0.01 | 0.005 | 0.001 |
|---|---|---|---|---|---|
| 1500 | 0.045 | 0.050 | 0.060 | 0.063 | 0.071 |
| 9900 | 0.017 | 0.019 | 0.023 | 0.025 | 0.028 |
| 18200 | 0.013 | 0.014 | 0.017 | 0.018 | 0.020 |
| 26600 | 0.011 | 0.012 | 0.014 | 0.015 | 0.017 |
| 35000 | 0.009 | 0.010 | 0.012 | 0.013 | 0.015 |

TABLE III
CRITICAL KS STATISTIC FOR DIFFERENT CONFIDENCE LEVELS.

**Numerical calculations.** To solve (5) for given $F_X$ and $F_Y$, i.e. to obtain the optimum KS distance $d_{KS}$ and corresponding scale and shift parameters $a^*$ and $b^*$ we used a gradient-free numerical optimization method based off the Nedler-Mead Simplex for a confined search space $(a, b) \in [1/1.5, 1.5] \times [-2, 2]$ (chosen after numerical experiments showed that most $a^*$ and $b^*$ values lie in this interval). Our implementation was in `C` with a `MEX` interface to `MATLAB`. Because of numerical errors, metricity (the number of triangles involving a particular point for which the triangle inequality holds) may not hold empirically for some real CDF pairs. However in our experiments this is not an issue, as the typical metricity is well over $90\%$.

### C. $d_{KS}$ is metric distance

**Lemma:** *The Kolmogrov-Smirnov Test under an affine transformation is a metric over the space of CDFs.*
**Proof:** A distance metric over a vector space must fulfill the following properties: non-negativity ($d_{KS}(F_X, F_Y) \geq 0, \forall \mathbf{X}, \mathbf{Y}$), identity of non-discernables ($d_{KS}(F_X, F_Y) = 0, \iff \mathbf{X} = \mathbf{Y}$), symmetry ($d_{KS}(F_X, F_Y) = d_{KS}(F_Y, F_X), \forall \mathbf{X}, \mathbf{Y}$; follows from properties of $p-$norms), and triangle inequality. The first three are rather trivial; we now show that the KS test 5 satisfies the triangle inequality. Let $F_1$, $F_2$ and $F_3$ be three cumulative distribution functions. Then $\exists\, a_2,\, b_2$ such that $d_{KS}(F_1,\, F_2) = |F_1(t) - F_2\left(\frac{t-b_2}{a_2}\right)|$, and $\exists\, a_3,\, b_3$ such that $d_{KS}(F_1,\, F_3) = |F_1(t) - F_3\left(\frac{t-b_2}{a_2}\right)|$. Then

$$\begin{aligned}
&d_{KS}(F_2,\, F_3) = \min_{a,b} \max_t \left| F_2(t) - F_3\left(\frac{t-b}{a}\right) \right| \\
&= \min_{a_2,b_2,a_3,b_3} \max_t \left| F_2\left(\frac{t-b_2}{a_2}\right) - F_3\left(\frac{t-b_3}{a_3}\right) \right| \\
&=\leq \max_t \left| F_2\left(\frac{t-b_2}{a_2}\right) - F_3\left(\frac{t-b_3}{a_3}\right) \right| \;\; \forall a_2,\, b_2,\, a_3, b_3 \\
&\leq \max_t \left| F_2\left(\frac{t-b_2}{a_2}\right) - F_1(t) \right| + \max_t \left| F_1(t) - F_3\left(\frac{t-b_3}{a_3}\right) \right| \\
&\leq \min_{a_2,b_2} \max_t \left| F_2\left(\frac{t-b_2}{a_2}\right) - F_1(t) \right| + \\
&+ \min_{a_3,b_3} \max_t \left| F_1(t) - F_3\left(\frac{t-b_3}{a_3}\right) \right| \\
&= d_{KS}(F_1,\, F_3) + d_{KS}(F_1,\, F_2)\, \square
\end{aligned} \tag{16}$$

### D. Multinomial Logit EDD class analysis

**Formulation.** We consider a set of $N$ users choosing to consume according to one of $K = 13$ EDD classes as extracted in Section IV. User $i$ chooses the shape $f$ as part of the decision process illustrated in Figure 4 (left), and bases his decision on his household-level characteristics $x_i$ (a selection of which are in Table I). In doing so, the user maximizes a utility function

$$U_{ik}(x_i) = V_{ik}(x_i) + \epsilon_{ik}. \tag{17}$$

We consider $V_{ik}(x_i) = x_i^T \beta_k$, where $\beta_k$ are choice-specific parameters to be estimated and $\epsilon_{ik} \sim Gumbel(0, 1)$. This is a typical formulation of a multinomial logit (MNL) model [5], which is well-understood in discrete-choice analysis. One can show that the probability of user $i$ to choose EDD class $k$ is

$$P_i(k) = \frac{e^{x_i^T \beta_k}}{\sum_h e^{x_i^T \beta_h}}. \tag{18}$$

The elasticity with respect to characteristic $t$ of the $i$-th user to choose EDD class $k$ is given in this model by $\zeta_{itk} = [1 - P_i(k)] x_{it} \beta_{kt}$, from which group elasticities are computed in each class $k$ by a weighted average (with weights $P_i(k)$):

$$\zeta_{tk} = \frac{\sum_h P_h(k) \zeta_{htk}}{\sum_h P_h(k)}. \tag{19}$$

**Estimation.** We estimate the model (17) using the standard maximum-likelihood framework (using the `multinom` package in `R`). Estimation was performed by setting Cluster 1 as reference. Because of the large number of covariates, we perform model selection using forward stepwise regression (using the `stepAIC` function in `R` that optimizes the *Akaike Information Criterion* [4], which is a measure of model fit). A relatively small number of characteristics (namely 17) enter with an appreciable and statistically-significant effect (under a Wald test at least $95\%$ level) in the decision of a EDD class (Table I). Moreover, the McFadden *pseudo-$\rho^2$* for estimation with the covariates in Table I is $0.17$, which is relatively low, but still comparable to other demand models in the literature [5]. When used for classification, the MNL reported here achieves a $26\%$ out-of-sample accuracy (in a 10-fold cross-validation setting), which may seem low, yet is much better than random guessing ($1/13 = 7\%$).