Drivers of variability in energy consumption

Adrian Albert[1], Timnit Gebru[1], Jerome Ku[1], Jungsuk Kwac[1], Jure Leskovec[2], Ram Rajagopal[3]

[1] Electrical Engineering Department, Stanford University
[2] Computer Science Department, Stanford University
[3] Civil and Environmental Engineering Department, Stanford University

Abstract. Meeting uncertain supply conditions with adequate demand-side measures is becoming increasingly central to the day-to-day operations of energy utility companies, as variability is one of the main drivers of cost in operating the grid. In this paper we provide the first characterization of demand variability and the factors influencing it. We propose that heterogeneity in energy consumption is driven by users' daily schedules, and mine a large sample of smart meter data from 30,000 residential users in CA ($\sim 17M$ daily load profiles) to extract a small *alphabet* of typical schedules that people follow throughout the day. We next relate variability in consumption to features pertaining to the individuals and their neighborhoods using data on customer engagement with the utility company and block-level U.S. Census demographics. Our analysis shows that certain high-level attributes - such as income level or tenancy situation - are robust predictors of schedule stability.

1 Introduction

In the large-scale infrastructure systems such as the power grid variability has a determining impact on cost-of-operation and environmental externalities. Thus, in designing practical interventions and efficiency programs at energy utility companies it is of relevance to identify the types of users or neighborhoods that contribute to variability on the grid, and to understand what changes in their their behavior would be most desirable for the system.

Electricity accounts for $\sim 40\%$ of total energy use and 34% of GHG emissions in the U.S. [1]. As intermittent renewable sources achieve higher rates of penetration, both demand and supply are becoming highly volatile. To better understand energy demand, utility companies in the U.S. have deployed millions of sensors (*smart meters*) capable of recording data at sub-hourly time resolutions. Profiling customer demand aids utilities in their market operations, e.g., with capacity planning and day-ahead scheduling. Moreover, understanding consumption decisions informs designing and implementing energy-efficiency and demand-response programs, and marketing to the appropriate user groups.

Here we provide the first study of energy consumption variability using smart meter data at two different levels of decision-making - the residential customer and their neighborhood. We mine a large dataset (17M days, 15-minute resolution) from 30,000 users in Northern CA to learn a small "alphabet" of daily

load profiles that encodes user lifestyle and schedules. Schedules represent the observed cumulative sum of latent usage choices; as such we hypothesize that the heterogeneity in schedule choice over time arises because of different attributes of users and the neighborhoods that they live in. We thus define a classification problem in which we use U.S. Census data on neighborhood demographics and user interaction with the utility company to predict schedule stability.

The rest of the paper is structured as follows. Section 2 introduces the variabiltiy analysis problem and the model of consumption used in this paper. Section 3 discusses the literature on smart meter data analytics. Section 4 describes our dataset. Section 5 discusses the algorithms that we use to extract such schedules from data. Section 6 presents a discussion of the drivers of variability in consumption schedules. We conclude in Section 7.

2 Problem statement

In Figure 1 we present a schematic of the methodology and analysis employed in this paper. We detail each of the components below.

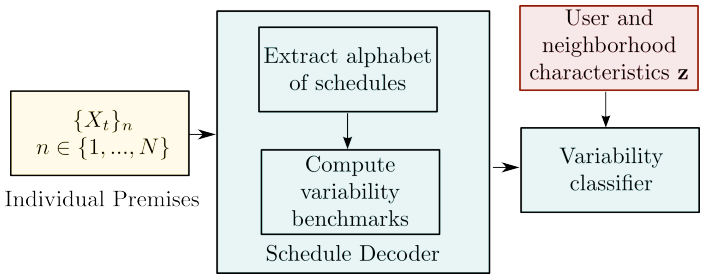

Fig. 1. Schematic of analysis methodology developed in this paper.

Daily schedule model. We observe (multivariate) series of consumption $\{\mathbf{X}_d\}_n$, $d = 1, ..., D$ (days) for individuals $n = 1, ..., N$. Following [2] and as described by each panel in Figure 2, we posit that the daily consumption profile $x_d(t)$ (also denoted by $\mathbf{x}_d$ in this paper) for a given user n and day d may be described by two quantities: *i)* the total daily consumption $E(d)$ and *ii)* the normalized daily schedule $s_d(t)$, with $t = 1, ..., T$, with $T = 4 \times 24 = 96$ (one measurement every 15 minutes for 24 hours in a day):

$$x_d(t) = E(d) \cdot s_d(t), \tag{1}$$

with $\sum_{t=1}^{T} s_d(t) = 1$. Since it is a measurement of physical energy, $X_d(t) \geqslant 0$, and the above amounts to taking the L_1 norm of $\mathbf{x}_d$, $E(d) = ||\mathbf{x}_d||_1$. Here we assume that the magnitude of consumption (the daily total $E(d)$) is determined primarily by factors exogenous to the user's schedule, i.e., weather or appliances. This paper focuses on studying the schedule component of consumption $s_d(t)$.

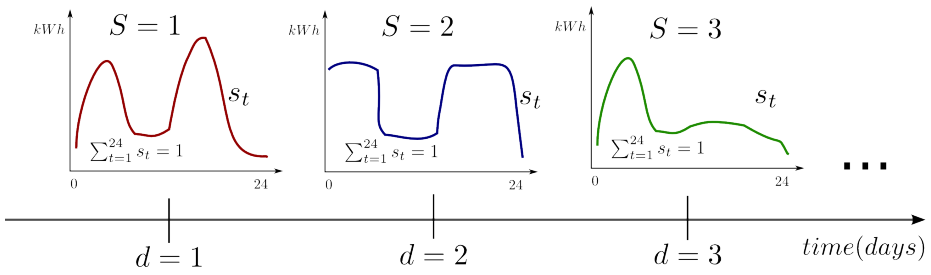

Fig. 2. Model of individual consumption composed of daily schedules chosen each day.

Alphabet of daily schedules. Using the set of all daily profiles recorded from the N users, we extract "typical" schedules $\hat{s}(t)$ using a clustering methodology. This step achieves a significant reduction of the data - millions of daily shapes may be described by using a common set of only a few tens of typical schedules.

Variability and stability benchmarks. One may use the alphabet of schedules extracted in the previous step to concisely describe the temporal sequences $S(d)$ of days in a user's consumption where each day assumes one value ("letter") from the alphabet. This situation is described in Figure 2, where a user's consumption schedules over three consecutive days $d = 1, 2, 3$ are $S(1) = 1$, $S(2) = 2$, and $S(3) = 3$. We quantify the variability (or conversely, stability) of user schedules using several benchmarks under simple assumptions of shape selection, both overall and by day-of-week. For a given user and model of schedule choice we compute the *entropy* over the alphabet sequences

$$S = -\sum_{\alpha} p(\alpha) \log p(\alpha), \tag{2}$$

and an upper bound $\Pi^{\max}$ on *predictability* (the maximum classification performance that the best algorithm may achieve) by solving an implicit equation

$$S = -\Pi^{\max}\log_2\Pi^{\max} - (1-\Pi^{\max})\log_2(1-\Pi^{\max}) + (1-\Pi^{\max})\log_2(M-1), \tag{3}$$

as done in [3], where M is the number of schedules required to describe the given user. The dependence of S with $\Pi^{\max}$ is illustrated in the left panel in Figure 4 for different sizes of the alphabet M. A given level of predictability (or stability) may be achieved by sequences of schedules of increasing entropy whose alphabet sizes are also increasing, albeit at different rates.

The models of schedule choice that we consider here are as follows:

1. *Completely at random*: the distribution over alphabet letters α is uniform, $p(\alpha) = \frac{1}{M}$. Here the entropy is computed as $S^{\text{rand}} = \log_2 M$.
2. *Independent and identically-distributed*: the choice of schedules is not correlated in time, but alphabet letters α are distributed *i.i.d* according to the frequencies of schedules observed in the dataset, $p(\alpha) = \frac{\#\{S(d)=\alpha\}}{D}$. The entropy in (2) may be computed as $S^{\text{uncorr}} = -\sum_{\alpha} p(\alpha) \log_2 p(\alpha)$.

3. *Serially-correlated*: the choice of schedules for the next day depends on past choices. Here the distribution over α in (2) is defined over sequences of schedules. We estimate the entropy in this case using the Lempel-Ziv algorithm [4] $S^{\text{full}} = \left(\frac{1}{D}\sum_d \Lambda_d\right)^{-1} \log_2 D$, with Λ_d the length of the longest subsequence α starting at day t which does not appear until time $d-1$.

Drivers of variability. We would like to understand whether certain information about the users and their communities may predict consumption variability. For each of the benchmarks computed above (either overall or by day-of-week) we separate the users into two classes of stability (either "low" or "high") based on a given benchmark quantile Q. We then learn a logistic classification model

$$y_n = h(\mathbf{z}_n^T \theta) + \epsilon_n, \tag{4}$$

where $y_n \in \{\text{Low}, \text{High}\}$, $h(\cdot)$ is the logistic link function, $\mathbf{z}_n$ is a vector of (standardized) characteristics for user n and their neighborhood, θ is a vector of coefficients to be estimated, and ϵ_n is a normally-distributed error term.

3 Literature review

Existing literature on analysis of smart meter data focuses on forecasting and load profiling applications. In [5] the authors use 15-minute resolution smart meter data from $\sim$ 200 customers of an utility company in Germany to group consumers according to their daily consumption profiles; they then develop different pricing schemes for each segment. Similar to our approach, in [6, 7] the authors describe intra-day consumption through a small number of recurring profiles. The K-means algorithm is by far the most popular statistical clustering approach. To populate a dictionary for representative shapes, the K-means algorithm can be a good starting point as used in e.g., [8]. In [9] the authors describe a two-stage pattern recognition of load curves based on various clustering methods including `K-Means`. [10] compares results obtained by using various clustering algorithms (hierarchical clustering, K-means, fuzzy K-means) to segment customers with similar electrical behavior. As an alternative approach to distance-based clustering, [11] introduces a class of mixture models, random effects mixture models, with a custom EM algorithm to fit the mixture models. More recently, [2] develops a customer segmentation methodology that is based on lifestyle patterns - typical load shapes identified using a `K-Means` algorithm.

4 Data description

The dataset that we use in this paper consists of an average of 48 months (January 2008 to December 2011) of consumption time series for 30,000 users in Northern California, along with information about some of their characteristics (such as education and general geographic location) and interactions with the utility company. From this dataset we selected 22,963 customers for which we

Table 1. Selected user descriptors

Owners	15003
Renters	7960
High School/Vocational	15374
College	5118
Graduate School	2471
Central Valley	7987
Coast	4515
Inland Hills	10307

Table 2. Selected block characteristics

	Min	1^{st} Q	Median	3^{rd} Q	Max
No. Days:	365	556	745	908	1397
Income($):	2499	46200	72650	92160	250000
Age:	14	31	37	43	81
Housing Value($):	9999	310800	494400	646700	1000000
No. Rooms:	2.0	4.4	5.2	6.0	9.0

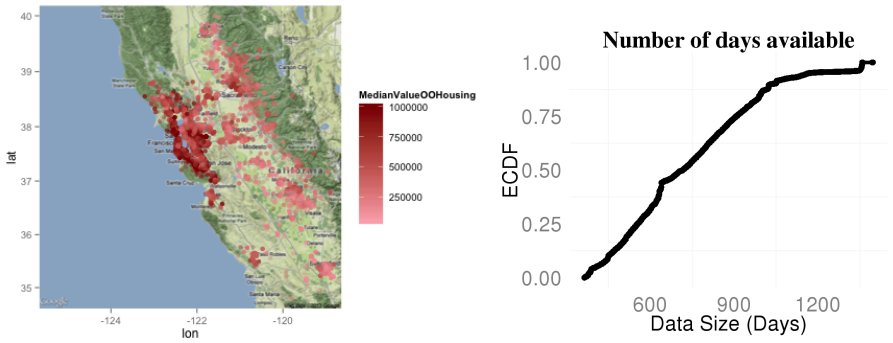

Fig. 3. *Left:* Geographic distribution of users colored by median housing value ($); *Right:* Distribution of data size (number of days available) across users.

had at least 1 year of energy usage data and who did not change residence during the selected time window. We further incorporated demographic data from the US 2006-2010 American Community Survey through its publicly accessible API.

Tables 1 and 2 provide a summary of selected categorical and numeric variables from user attributes and block-level demographics data used in our analysis. In addition, in Figure 3 (left panel) we present the geographic distribution of users color-coded by median housing value. In the right panel in the figure we present the distribution of data size (number of days available) across users.

5 Clustering daily load profiles

`K-Means` clustering. We would like to cluster the input multivariate time series $\{\mathbf{X}_d\}_{n=1,...,N}$ (after scaling by the daily total energy $E(d)$) to obtain an alphabet $\mathcal{S} = \{\hat{s}^k(t)\},\ k \in 1, ..., K$ of K typical schedules. By far the most popular "general-purpose" clustering algorithm is `K-Means` [12], which typically uses an Euclidean distance between elements in the cluster

$$d(\mathbf{x}, \mathbf{y}) = ||x - y||_2, \tag{5}$$

where $\mathbf{x}$ and $\mathbf{y}$ are daily profiles as in Figure 2, i.e., vectors in $\mathbb{R}^T$, with $T = 96$. The `K-Means` algorithm is an iterative procedure to find a local solution to the optimization problem

$$\min \sum_{k} \sum_{x_j \in \mathcal{C}_{\|}} d(\mathbf{x}_j, \mathbf{c}_k)^2, \tag{6}$$

where $\mathbf{c}_k \in \mathbb{R}^T$ is the center profile of cluster k.

Drawbacks of `K-Means`. For smart grid applications such as forecasting and control it is essential to understand the formation of *peaks*, as they contribute to much of the environmental and financial costs of energy consumption. In this context, both timing of peaks and pace of ramping up (or down) are key concepts - consuming 1 kWh more during 5 minutes in the peak-time at 5 PM is much more expensive (and polluting) than consuming the same 1kWh over one hour during the night at 3AM, when aggregate demand is generally low. However the L_2 norm penalizes uniformly the mismatch across time-of-day, and is agnostic of ramping. In particular, this measure applies a "double-penalty" to profiles that are only slightly different in timing when peaks occur, such as in the situation in Figure 4 (right panel). There, profiles $S1$ (red line) and $S2$ (green line) have each a 30-minute peak that is separated by one hour, while profile $S3$ (blue line) is flat. Computing the Euclidean distance between the three profiles we have $d(S1, S2) = 0.006$ and $d(S1, S3) = 0.003$. Thus, in a clustering application, `K-Means` using the L_2 norm would rather assign the flat profile $S3$ in the same cluster with $S1$, when in fact $S1$ and $S2$ are similar.

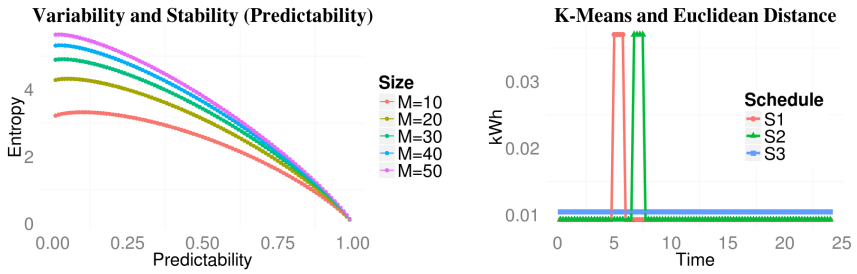

Fig. 4. *Left:* Benchmarks for variability (entropy) and stability (predictability bound [3]) for different alphabet sizes N; *Right:* Sample energy profiles that illustrate the drawbacks of the Euclidean distance with `K-Means`.

Accounting for shifting and scaling. A partial solution to the issues raised above is offered by the *K-Spectral Centroids* (`KSC`) algorithm [13]. The algorithm defines a distance metric between two profiles $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\hat{d}(\mathbf{x}, \mathbf{y}) = \min_{\alpha, q} \frac{||\mathbf{x} - \alpha \mathbf{y}_q||}{||\mathbf{x}||}, \tag{7}$$

where α is a scaling parameter and $\mathbf{y}_q$ is a shifted version of $\mathbf{y}$ by an amount q that minimizes the distance $\hat{d}$. In our implementation we allow for an hour and 15 minute shift which means that q can be between -5 and 5. The distance calculation first minimizes (7) with respect to q using a grid-search procedure; once q is determined the optimal scaling α^* can be found by setting the gradient to zero since $\hat{d}$ is a convex function of $\mathbf{y}_q$, yielding $\alpha^* = \frac{\mathbf{x}^T\mathbf{y}_q}{||\mathbf{y}_q||^2}$. With this new distance metric, a cluster center is now defined to be the centroid μ_k that minimizes $\mu_k^* = \arg\min \sum_{x_i \in C_k} \hat{d}(x_i, \mu)^2$. It is shown in [13] that the solution to this optimization problem is given by

$$\mu_k^* = \arg\min \frac{\mu^T \mathbf{M} \mu}{||\mu||^2}, \tag{8}$$

where $\mathbf{M}$ is a matrix where each row consists of a daily profile in cluster k. The solution to this minimization problem is the eigenvector μ corresponding to the smallest eigenvalue λ_M of matrix $\mathbf{M}$.

Choosing alphabet size. Finding the appropriate value of the number of clusters K in the algorithms above is generally considered an open problem in machine learning; many techniques have been proposed for that purpose such as the *Average Sihlouette*, the *Gap Statistic*, or *Hartigan's Index* [12], [13] that make different assumptions about the clusters sought. Here we use a much simpler criterion - we would like to obtain an alphabet of schedules that covers at least 50% of the variance in the data, defined as Coverage = $\frac{\text{Between Sum of Squares}}{\text{Total Sum of Squares}}$. In our experiments, a value $K = 50$ achieves this goal. When comparing the sum of squared distances between the cluster centers (a popular measure of cluster separation, the higher the better) for the results obtained with the two algorithms we obtain 929.98 for `KSC` and only 56.14 for `K-Means`, which indicates that `KSC` does indeed produce more diverse cluster centers.

6 Variability and its drivers

6.1 An alphabet of schedules

Figure 5 compares 50 shapes that are obtained after clustering the entire consumption data using `KSC` as discussed above. We implemented `KSC` in Java on Amazon's Elastic MapReduce (EMR) service. The algorithm converged after 30 iterations for $K = 50$, and the schedules obtained cover at least 50% of the variance in the data. The most frequent schedules (C6 and C3) display consumption activity clustered in the evening. Double-peak schedules (such as C34, C49, C14, or C17) - which is the default view of consumption at utility companies - cumulatively account for only 15% of all consumption. This finding is consistent with literature values (on much smaller data) [8]. Interestingly, there are quite a few schedules with pronounced activity mid-day (e.g., C33, C8, C32, C44), which cumulatively account for about 11% of all consumption.

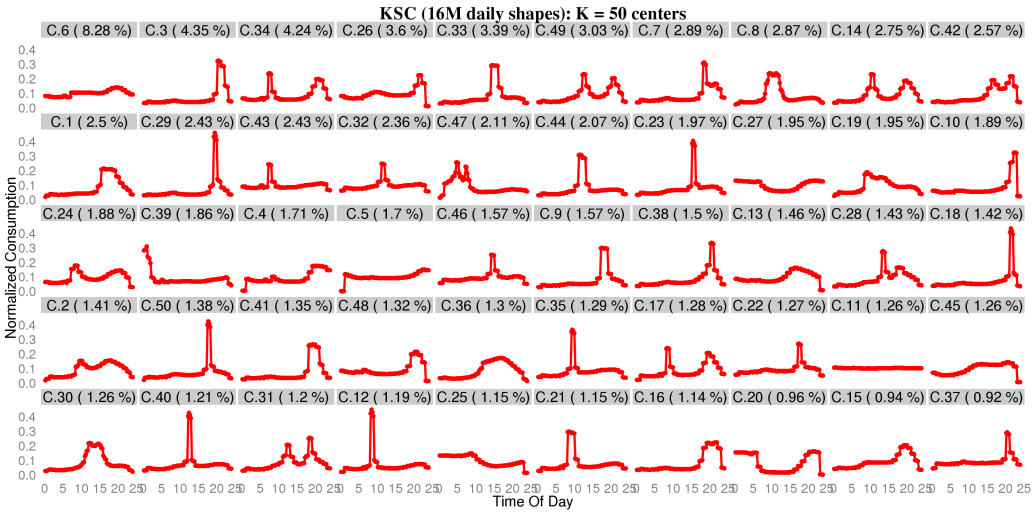

Fig. 5. 50 centers obtained using `KSC` (ordered by support in the data).

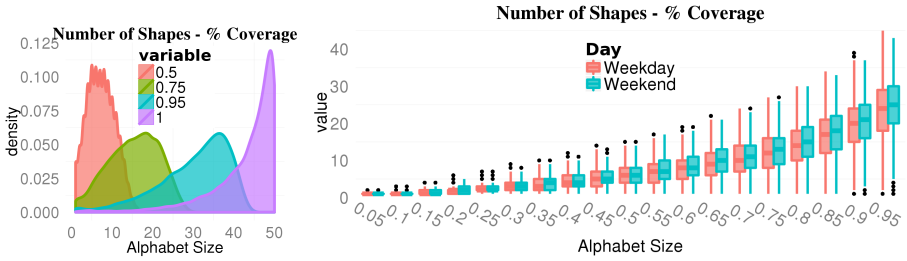

Fig. 6. *Left:* Population distribution of the number of typical schedules needed to achieve a target coverage (50%, 75%, 95%, 100%) of individual daily consumption.

6.2 Benchmarks of stability

Using the schedules alphabet extracted above we investigated the size of the individual subsets of the alphabet in Figure 5 needed to achieve certain values of coverage of schedule sequences. We present the calculations in Figure 6. The distribution plots in the left panel suggest that users will typically follow a restricted subset of schedules most of the time - only ~ 15 schedules are enough to cover 75% of days for most people. Yet there is a difference of about ~ 15 schedules going from 95% to 100% coverage, which indicates that "unusual" schedules are quite different among each other. In the left panel we present the dependence of the average alphabet size required for given levels of coverage broken down by weendays and weekends. We observe that weekends will consistently require a larger number of schedules than weekdays for the same level of coverage, which

is again not surprising - presumably people engage in more diverse activities during days off than during work days.

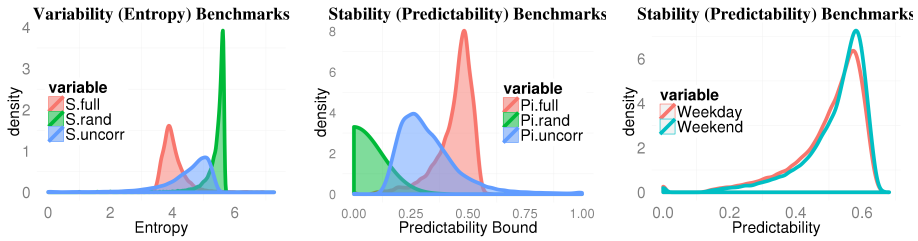

Fig. 7. *Left:* Entropy benchmarks for three models of schedule selection; *Middle:* predictability (stability) benchmarks; *Right:* Distribution of $\Pi^{\max}$ (under the serial correlations model) by weekend/weekday.

We next computed entropy and stability (predictability) benchmarks as in Section 2 for each of the 22,963 users in our final sample. In Figure 7 we present the distribution of entropy S (left panel) and predictability bound $\Pi^{\max}$ (right panel) over the entire sample for the three models of schedule choice outlined in Section 2. In the left panel, from left to right, the density curves are S^{rand} (uniformly at random schedule selection model), S^{uncorr} (i.i.d. schedule selection model) and S^{full} (the full entropy estimated from sequences of schedules as in [3]). This is a clear indication that consumption at the individual level is not purely random, since in general $S^{\text{uncorr}} < S^{\text{rand}}$. Moreover, schedule decisions in the present depend on decisions in the past (i.e., there is information in the temporal sequences of schedules), since overwhelmingly $S^{\text{full}} < S^{\text{uncorr}}$.

Computing $\Pi^{\max}$ (middle panel in the figure) we observe, as expected, that predictability is highest when accounting for temporal correlations in daily schedule selection across time. However for most individuals we have $\Pi^{\max} \sim 55\%$, i.e., we may hope to correctly predict the type of schedule followed by the typical person only around 55% of the time. This suggests that even if there are in general only a small number of typical schedules in our alphabet $\mathcal{S}$, consumption is still volatile because of heterogeneity in how users select among these daily schedules over time. In turn, this suggests that attempting to forecast daily schedules for individuals may be inherently difficult, and motivates our future work on understading how widely variable schedules aggregate to stable averages for even small, geographically-homogenous groups. A surprising observation can be made looking at the right panel in Figure 7: while consumption on weekends follows a larger set of schedules as discussed above, it is on average at least as predictable (or more) than that on weekends. This indicates that temporal correlations do play a role in determining choice of weekend consumption.

6.3 Drivers of stability

We learned a logistic regression model (4) that classified users into low or high classes on the predictability benchmark. Significant regression coefficients (at least at the 0.05 level on the t-test) are presented in Figure 8 for user-level attributes (left panel) and neighborhood-level characteristics (right panel) for a $\{\text{Low}, \text{High}\}$ breakdown corresponding to $Q = 0.5$ (median).

Among the most important drivers of schedule stability is the climate zone - in particular the hot zone W (Central Coast). Users in this climate zone will likely have a large AC (cooling) component in their consumption, which may give rise to regular afternoon consumption. Perhaps unsurprisingly, when people rent their consumption is less predictable - e.g., because owners will tend to be more responsible with their consumption when they pay for their own utility bills. Engagement with the utility (as indicated by the number of interactions or complaints - "tickets" generated) does correlate with users being more predictable. Education plays a small role, too - the less educated users (having only a high-school degree) will tend to consume less predictably. Whether the user has applied for efficiency-motivated rebates ("Appliances", "HVAC", "Lighting" etc. in Figure 8) - which again may be interpreted as indication of interest in energy use - does have a moderate, but significant effect on schedule predictability. Yet we note that frequent interaction with the utility (number of "tickets") is a much stronger indicator of stability.

The most important neighborhood-level driver of stability is the median house value - the richer the neighborhood, the more predictable schedules its inhabitants follow. This enforces the observation before about the effect of education - in the Bay Area (as in may other regions in the U.S.) typically more educated people will tend to live in richer neighborhoods that have higher property values. The next important predictor is house size - houses with 5 bedrooms and more will have more predictable consumption. This is surprising - rules of thumb used in practice suggest that more occupants will yield a more volatile aggregate consumption. Similarly, the more likely the house is to use natural gas as heating source, the more predictable its energy use will be. Understandably, the more people pay for rent (either in absolute terms or as fraction of income), the more stable their consumption will be - presumably because they use fewer appliances that contribute to volatility.

An illustration of the robustness of the regression results is summarized in Figure 9. The left panel illustrates the Receiver Operating Characteristic (ROC) curve for the logistic regression used as classifier (the tuning parameter is the probability threshold separating the two classes). We show curves for the overall stability estimates, as well as for Tuesdays and Saturdays. All classifiers achieve better than random performance since the curves are above the diagonal line in the ROC space. Interestingly, the classifier performs better (larger area under curve) on full sequences (as opposed to on a by-day basis), which indicates that the rather general characteristics we employed do not hold enough discriminative power to resolve finer distinctions such as weekday/weekend stability.

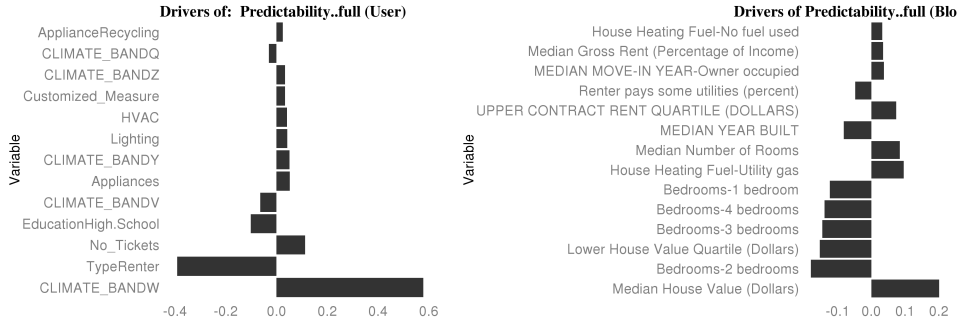

Fig. 8. Regression results: user (left) and neighborhood features (right).

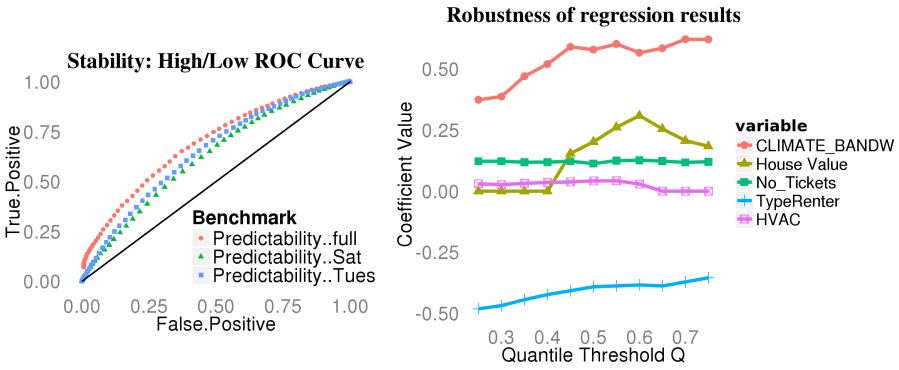

Fig. 9. Classification robustness. *Left:* ROC curves for full series, Tuesdays, and Saturdays; *Right:* Robustness of coefficient estimates for different thresholds Q of stability.

In the right panel we show the behavior of coefficient estimates for several selected variables (with high contributions at $Q = 0.5$) as the definition of the $\{\text{Low}, \text{High}\}$ stability classes is changed by varying Q from 0.25 to 0.75 (the inter-quartile range of $\Pi^{\max}$). Note that most variables maintain their relative magnitude relationship, which indicates that strong effects are generally well resolved. `House Value` gains in importance as the high predictability class contains more of the more stable users.

7 Conclusions

We have developed a methodology for analyzing variability in energy consumption based on identifying recurring patterns in users' daily schedules. We learned an alphabet of shapes from a large sample of 22,963 CA users comprised of $\sim 17M$ daily load profiles and used it to characterize choice of schedules. Using demographic data, we have identified key neighborhood and individual charac-

teristics that determine consumers' predictability. We are currently developing a more natural clustering algorithm for consumption profiles, as well as a parametric model of schedule selection that incorporates temporal correlations.

Acknowledgements

We are grateful to Chin Woo Tan and June Flora for helpful discussions related to operational planning at energy utility companies. We thank Brian Smith and Ann George at PG&E for providing the dataset.

References

1. Energy Information Administration. Annual energy review. Report DOE/EIA 0384, U.S. Department of Energy Office of Energy Markets and End Use, Washington, DC, 2007.
2. Jungsuk Kwac, June Flora, and Ram Rajagopal. Household energy consumption lifestyle segmentation using hourly data. *To appear*, 2013.
3. Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-Lszl Barabsi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
4. T. Schurmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414–427, 1996.
5. Christoph Flath, David Nicolay, Tobias Conte, Clemens van Dinther, and Lilia Filipova-Neumann. Cluster analysis of smart metering data - an implementation in practice. *Business & Information Systems Engineering*, 4(1), 2012.
6. Teemu Rasanen and Mikko Kolehmainen. Feature-based clustering for electricity use time series data. In Mikko Kolehmainen, Pekka Toivanen, and Bartlomiej Beliczynski, editors, *Adaptive and Natural Computing Algorithms*, volume 5495 of *Lecture Notes in Computer Science*, pages 401–412. Springer Berlin / Heidelberg, 2009.
7. 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.
8. Brian Artur Smith, Jeffrey Wong, and Ram Rajagopal. A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting. *ACEEE Summer Study on Energy Efficiency in Buildings*, 2012.
9. G. Tsekouras, N. Hatziargyriou, and E. Dialynas. Two-stage pattern recognition of load curves for classification of electricity customers. *IEEE Transactions on Power Systems*, 22:1120–1128, 2007.
10. 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.
11. G. Coke and M. Tsao. Random effects mixture models for clustering electrical load series. *Journal of Time Series Analysis*, 31(6):451–464, 2010.
12. T. Hastie, R. Tibshirani, and J.H. Friedman. *The elements of statistical learning: data mining, inference, and prediction.* Springer series in statistics. Springer, 2009.
13. Jaewon Yang and Jure Leskovec. Patterns of temporal variation in online media. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 177–186, New York, NY, USA, 2011. ACM.