Poster Abstract: Segmenting Consumers Using Smart Meter Data

Adrian Albert
Electrical Engineering
Stanford University
adalbert@stanford.edu

Ram Rajagopal
Civil and Environmental Engineering
Stanford University
ramr@stanford.edu

Raffi Sevlian
Electrical Engineering
Stanford University
raffisev@stanford.edu

Abstract

Existing electricity market segmentation analysis techniques only make use of limited consumption statistics (usually averages and variances). In this paper we use power demand distributions (PDDs) obtained from fine-grain smart meter data to perform market segmentation based on distributional clustering. We apply this approach to mining 8 months of readings from about 1000 US Google employees.

1 Introduction

Detailed analytics of fine-grained *smart-meter* data is expected to both increase the energy efficiency and performance of the electricity market and mitigate carbon emissions through more efficient generation [4]. Demand-response programs may address the stress that individual usage choices place on generators, while utilities may better forecast the amount of electricity they need to buy based on the demand patterns of consumers groups [1].

Currently, market segmentation is performed using coarse consumption data and statistics (e.g., [5]). We propose a market segmentation approach that is based on distributional clustering of PDDs, which *i)* involves a statistically meaningful distributional distance; and *ii)* is invariant to scaling and shifting in the power demand.

To the best of our knowledge no study of similar scope has been proposed to date. The closest match to ours is [6] from the computer science literature, where popularity profiles in online media are clustered according to their shapes.

2 Patterns of power demand distributions

Given N electricity consumption time series we build discrete empirical cumulative distribution functions (CDF) $F_X(x)$. We are interested in identifying a small number K ($K \ll N$) of groups (PDD *classes*) such that the distributions are statistically similar within groups and statistically different across groups. Groups represent segments of customers that consume electricity in markedly different ways.

PDD classes. Two users with PDDs specified by the underlying random variables $\mathbf{X}$ and $\mathbf{Y}$ shall be termed similar if

$$\mathbf{Y} = a\mathbf{X} + b, \tag{1}$$

where a is the *scale parameter* and b is the *shift parameter*. It is easily shown that the transformed CDF will be given by $F_{Y;a,b}(y) = F_X(y^*)$, where $y^* = (y-b)/a$. Users $\mathbf{X}$ and $\mathbf{Y}$ have similar PDDs (but different means and variances).

Distributional similarity. To compare two distributions F_X and F_Y we use the well-known two-sample *Kolmogorov-Smirnoff (K-S) statistic* $d_{KS}(F_X, F_Y) = \max_x |F_X(x) - F_Y(x)|$. In particular, we wish to assess the optimum distributional distance between a distribution F_X and a distribution $F_{Y;a,b}$ obtained through (1):

$$d_{KS;a^*,b^*}(F_X, F_Y) = \min_{a,b} \max_x |F_X(x) - F_Y(y^*)| \tag{2}$$

We show in the extended version of this paper that $d_{KS;a^*,b^*}$ is indeed a distance metric over the space of CDFs.

PDDs and the electricity market. PDDs are important because they reflect the true cost of service to the utility. Utilities procure energy in advance from the electricity markets to satisfy demand in aggregate. Electricity is cheaper for less variable aggregate demands. But, the cost of variability 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 differentiated pricing and demand response programs offered by utilities. One can show [1] that the expected cost to the utility to serve a consumer whose consumption $\mathbf{L}$ (CDF $F_L(s^*)$) is $C^* = p\,\mathbb{E}[\mathbf{L}|\mathbf{L} \geq s^*]$ [\$/hr]. Here the utility procures energy s [MW] in the day ahead at a cost p [\$/MW].

3 Mining PDDs from data

The K-Medoids algorithm. We cluster PDDs using a variant of the *K-Medoids* algorithm [2]. We chose this approach over the popular *K-Means* algorithm to avoid averaging for computing a cluster centroid, as this operation is not well-defined in the case of distributions. We initialize the clusters by fitting Gaussian mixture models to the distributions. We choose the number of clusters by studying the Average Silhouette heuristic [2] and obtain $K^* = 24$ for the total distributions, and $K^* = 40$ for peak/offpeak.

Computing distributional distances. To approximate the (generally non-convex) outer objective in (2) we developed

BuildSys'11, November 1, 2011, Seattle, WA, USA.
ACM 978-1-4503-0749-9

several heuristic approaches that first obtain a fast, coarse estimate of the solution, which is then used to initialize standard simplex-based or gradient-based optimization methods. To compute (2) for given a and b one can perform interpolations (in $O(L)$ time) to obtain CDF values for those shifted and scaled points in the support for which actual values are not available due to limited CDF resolution. To eliminate the need for interpolation we developed an approach in which we represent each CDF by a small set (of size $W \ll L$) of piecewise quadratic polynomials for which the scale and shift transformations can be obtained in closed-form in $O(W)$ time. For our initial results presented here we could achieve a $x8 - x10$ speed-up for $L = 1000$ bins. Details of the aforementioned heuristics and distance acceleration algorithm will be given in a forthcoming paper.

4 Experimental setup and results

Dataset description. The data used in this study was composed of 10-minute smart meter readings electricity consumption time series of 949 (1,112 before data clean-up) US-based Google employees households located mainly in Silicon Valley. This dataset (*Surge*) was collected by Stanford and Google researchers in March-October 2010 [3]. To compute discrete *pdfs* and *CDFs* we defined a grid of $L = 1000$ bins from 0...2.5 $kWh/10$ minutes and used the Amazon EC2 MapReduce framework for mining large datasets.

Overall PDD examples. As shown in Figure 1 (top-left), obtained clusters have average inter-cluster K-S distances (blue line) of less or about 0.05, which represents an acceptance of the null hypothesis that the distributions they contain are similar at a 5% confidence level under the two-sample Kolmogorov-Smirnoff test and model (1). The average inter-cluster distance (top-middle panel, green line) is consistently higher than the intra-cluster distance. The largest intra-cluster distance is achieved for Cluster 21, which contains multi-peaked distributions: more shape variability tests the limits of the model (1). Example cluster centers (blue: pdf, green: CDF) are given in Figure 1 (top-right).

The bottom-left panel presents a heat map plot the cluster members corresponding to those centers. The clustering will identify general patterns in the consumption distributions that are qualitatively different from each other. Users in Cluster 1 have sharp, single-peaked, short-tailed power demand profiles, which indicates a preference to low-intensity energy usage. Their low-variance demand pattern imposes relatively low servicing cost on the utility. In contrast, Cluster 21 users have stretched, multi-peaked power demand profiles, which indicates more variability in consumption.

Peak and off-peak demand. For differential pricing it is useful to contrast user demand behavior in the peak time (7:00AM-8:00AM and 4:00PM-8:00PM) with that in the off-peak. For example, we test the hypothesis that some clusters are robust to the offpeak-peak transition, i.e., most of the users in one such cluster will also display consistent behavior during peak time (they will end up grouped in a common cluster). The lower-right panel in Figure 1 presents the degree of consistency among user behaviors under the offpeak-peak time transition. We have investigated the percentage of users from each offpeak cluster that are allocated to a common cluster in the peak time (blue line), and the percentage of peak-time users that come from a common cluster in the offpeak (green line). For example, for Cluster 8, 40% membership is preserved, and for Cluster 28 60% membership is preserved. Moreover, for Cluster 28 in the peak time (the multi-peaked cluster), all members come from Cluster 28 in the offpeak time (100% consistency).

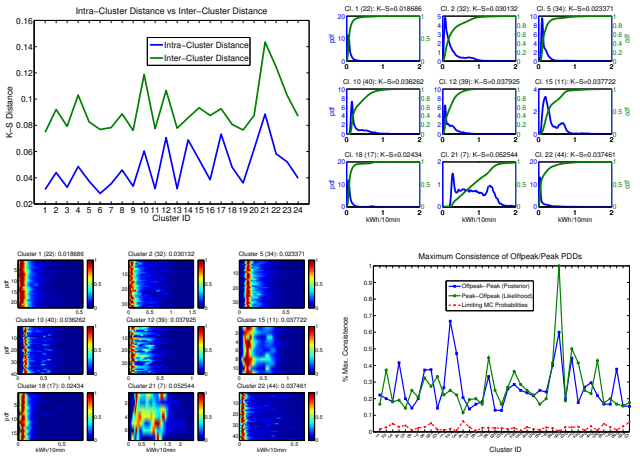

Figure 1. ***Top row*****. Left: inter- and intra-cluster distance for the clustering solution. Right: 9 example cluster centers pdfs (blue) and cdfs (green).** ***Bottom row*****. Left: heatmap of cluster member pdfs (each row color-codes a distribution). Right: demand behavior consistency for offpeak (blue line) and peak (green line) clusters.**

5 Future work

We are currently working on a detailed time analysis (time-of-day and day-of-week) of PDD clusters. Using survey data on the households, we are investigating the effect of socio-economic indicators (e.g., income, age, gender, location) on consumption behavior. We are also developing a model of choice of consumption behavior to inform the design of demand-response and energy policy analysis.

6 References

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

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

[3] S. Houde, A. Todd, A. Sudarshan, J. Flora, and K. C. Armel. Real-time feedback and electricity consumption: a field experiment assessing the potential for savings and persistence. *Submitted*, July 2011.

[4] A. Krioukov, C. Goebel, S. Alspaugh, Y. Chen, D. E. Culler, and R. H. Katz. Integrating renewable energy using data analytics systems: Challenges and opportunities. *IEEE Data Eng. Bull.*, 34(1):3–11, 2011.

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

[6] J. Yang and J. 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.