Finding the right consumers for thermal demand-response: an experimental evaluation

Adrian Albert[1] and Ram Rajagopal[2]

***Abstract*—For demand-side management programs concerned with heating, ventilation, and air conditioning (HVAC) energy consumption, smart meter data collected at the whole-premise level has recently been used to decompose usage into its HVAC and non-thermal components, which are typically not separately monitored. In this paper, we study the extent to which program design and decisions based on models using whole-home energy consumption differ from decisions made with full knowledge of appliance-level end-use patterns. We develop a model assessment methodology for the case when model results are used to rank consumers by their potential for thermal demand-response. For this, we compare rankings of consumers in two scenarios–first when only the aggregate outcome of the top consumers matters, then when the relative ordering of the consumers is important. We illustrate our methodology using two individual consumption models that extract *thermal* (temperature-sensitive) and *occupant-driven* components from single-point source smart meter data. Moreover, we discuss how a demand response program that selects the consumers with the most potential for energy reduction based on model results may achieve similar results as in the ideal case when separately-monitored HVAC data is used.**

***Index Terms*—HVAC control, smart meter data, consumer ranking, thermal response estimation**

NOMENCLATURE

S_n, $\hat{S}_n$	Estimated total consumption averted for the top n consumers ranked by either thermal regimes estimates or by the ground truth thermal response.
Δ_n, E_n	Absolute and relative mismatch between S_n and $\hat{S}_n$
λ	Threshold indicating how large an estimated value (a_t or σ_t) should be in comparison to its minimum value
$\mathbf{r}$	Ranking of consumers according to estimated thermal response
$\mathbf{r}$, $\hat{\mathbf{r}}$	Rankings of consumers by either model-estimated or ground truth thermal response rates
$\{x_t^i, T_t^i\}_{t=1}^{\mathcal{T}^i}$	Respectively, the observed series of consumption (in kWh) and temperature (in °F) for customer i at time t, across a horizon $\mathcal{T}^i$
$a^i(T)$, $\hat{a}^i(T)$	Thermal regimes model estimate and ground truth thermal response rates for consumer i at a temperature level T
a_j, b_j, σ_j	Respectively, thermal response rate (in kW/°F), base load (in kWh), and activity-induced consumption volatility (in kWh) for thermal regime j
c_{jl}, d_{jl}	The base utility and, respectively, the rate of change with temperature from transitioning from regime j to regime l
$d_{\text{rank}}(\mathbf{r}, \hat{\mathbf{r}}; T, n)$	Rank distance between rankings $\mathbf{r}$ and $\hat{\mathbf{r}}$ around temperature level T of the top n consumers
o_t^i, h_t^i	Separately-monitored consumption activity and, respectively, HVAC usage data for customer i at time t (in kWh)
$q_{a,\gamma}$	The $\gamma - th$ quantile of the distribution of a

[1]C3 Energy, Inc. (Work done while at Stanford University)
[2]Stanford University, Civil and Environmental Engineering Department.

I. INTRODUCTION

To design effective energy demand-side management (DSM) programs, it is key to understand which consumers may be beneficial to enroll, and to target them with appropriate controls. To this end, smart meter data offers unprecedented insight into how residential customers use energy throughout the day, as it is collected at high granularity (i.e., sub-hourly) and reliability from each customer. However, actionability on this data is limited, since by itself, it does not provide an understanding of how energy is consumed across different end-uses, particularly large ones like HVAC. As such, recent work has focused on decomposing smart-meter-monitored usage into its HVAC and non-thermal components, which are, typically, not separately monitored. This invites the question of how to compare the quality of the decisions based on model estimates to those taken if the complete information were available.

As pointed out recently in [1], it is often the case when evaluating machine learning algorithms that the end-goal—which is to inform decision-making—is lost to discussions of performance metrics that do not explicitly internalize the practical application for which the algorithm was developed. Understandably, researchers want to know how well their model or algorithm does compared to "ground truth". However, in this paper, we argue that in some applications the appropriate question to ask is rather how similar the outcomes are of decisions made based on the model with the decisions made when full information is available. Our case of interest is using estimated thermal response rates (i.e., the rate of change in energy consumption with a change in temperature) to rank consumers according to their potential for offering demand-response flexibility. Here, enrollment decisions are based either on models using whole-home consumption, or full knowledge of appliance-level end-use patterns.

The contribution that this paper makes is the methodology to compare the outcomes of using consumption models as opposed to ground truth HVAC data for the purpose of a demand response program. This program requires consumers to change their thermostat setpoint, thereby achieving energy reductions. The proposed methodology is summarized in Fig. 1. We are interested in two cases: in the first scenario, only the aggregate outcome of the top consumers matters; in the second scenario, the relative ordering of the consumers is important. For each consumer, we observe both hourly whole-home data and, importantly, recordings of usage over time

separately for a number of appliances. These end uses are both major consumers of electricity (e.g., HVAC) that may be either automated or may be influenced by user actions, and small loads (e.g., toasters, ovens) activated by consumer activity. To describe HVAC and activity-related consumption, we compare two models, a simple linear "breakpoint" model [2] and a model we have developed previously [3] that decomposes single-point source smart meter data into *thermal*, *baseload*, and *occupant-driven* components. We rank and compare consumers according to the amount of potentially averted energy by estimating the responsiveness of consumption to outside temperature. Then, we quantify the matching in the rankings obtained by either applying a statistical model or by employing the ground truth HVAC temperature sensitivity. In our analysis we assume well-intentioned residential energy consumers, i.e., no meter tampering or strategic behavior.

A. Related literature

Thermal consumption is an important part ($\sim 25\%$) of the residential energy budget in the U.S.; as such, much research has been dedicated to modeling the impact of weather (particularly temperature) on residential energy use. Prior research [4], [5] has indicated the substantial potential of using residential consumer flexibility in HVAC consumption for participation in demand-response programs.

While past studies had access mostly to aggregate (monthly) data (e.g., [6], [7]), newer studies such as [8] uses experimental and smart meter data to describe intra-day effects. In [9], a method is proposed to decompose whole-home usage into four categories (i.e., base load, activity load, heating season gradient, and cooling season gradient) using a breakpoint regression model. Other studies [10] model the HVAC operation starting from physical principles. The demand-response potential of thermostatically-controlled appliances such as HVAC is quantified in [11]. Most approaches develop a statistical model of individual consumption that separates thermal and non-thermal consumption. However these types of models are typically static—they do not allow for different temperature sensitivities based on temperature, or on different levels of user activity throughout time.

In [3], we proposed a simple semi-supervised thermal decomposition algorithm based on a hidden Markov model that interprets observed smart meter data as the consequence of unobserved decisions of the user to use or not to use HVAC (for either heating or cooling). In [12], we also show that patterns extracted from such a model are predictive of both the presence of large appliances and of user lifestyle attributes. A related problem is that of energy disaggregation, whereby individual appliance signals are recovered from whole-home consumption data without the need of additional instrumentation. One of the most cited uses of disaggregated appliance information is offering consumers detailed feedback about what appliances (or type of appliances) are being used at a given point in time, and what impact each end-use has to the overall energy bill [13]. E.g., in [14], the authors attempt to uncover many typical end uses through very granular (Hz-range) consumption.

Most literature on this subject focuses, however, on detecting the appliances that are being from whole-home consumption data. We could not find a single study in recent literature where the disaggregated estimates of HVAC and occupant-induced consumption were used to design demand-side management programs. This paper addresses this gap.

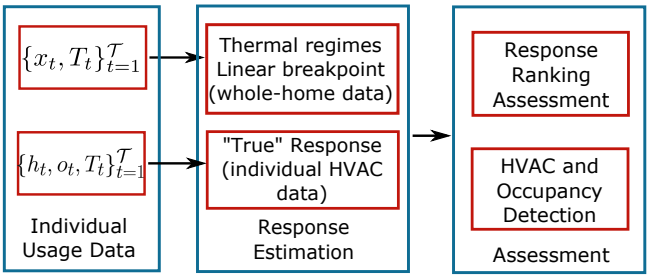

Fig. 1. Methodological outline for the experimental evaluation.

The structure of this paper follows the outline in Fig. 1. In Section II, we briefly describe the statistical models we use to describe individual HVAC response using either whole-home consumption data or separately-monitored HVAC data. In Section III, we describe the methodology used to assess how these models can be used to rank consumers for demand-response, and to detect significant occupant activity. In Section IV, we describe the experimental data. In Section V we apply the methodology to the data. We conclude in Section VI.

II. Individual Consumption Model Review

Data. In our experiment, we collect data from N individual households. For each household i, $i = 1, ..., N$, and at each hour[1] t, with $t = 1, ..., \mathcal{T}^i$ (where the $\mathcal{T}^i$s are time horizons that range from one to three years), we observe time series of total energy consumption x_t^i, of weather variables, out of which we focus on temperature T_t (which we assume is the same for all consumers in a homogeneous geographical region), and, importantly, of *separately-monitored* appliances usage. We further aggregate some of the separately-monitored appliance data into three broad categories, as discussed in Section IV: *thermal* h_t^i (HVAC), *base load*, and *occupant-induced* (intentional) o_t^i (non-HVAC)[2]. These are all inputs to the ranking methodology, as shown in Fig. 1. We make the (benign) assumption that there is no missing meter data.

We use this data to learn separate, individual models for every consumer i, in particular being interested to estimate "thermal response rates" of consumption with temperature, which we denote by the symbol a. We compare two models that use whole-premise data against the separately monitored HVAC data. Below, we briefly review the models mainly to fix notation and intuition, dropping the superscript i for simplicity.

The linear breakpoint model. A recent, popular approach to modeling the temperature response of residential consumers is the linear breakpoint model in [9]. In this model, observed whole-home readings x_t are assumed i.i.d and linear with T_t:

$$x_t = x_0 + a_-(T_t - T_0)_- + a_+(T_t - T_0)_+ + \epsilon_t, \qquad (1)$$

[1]Originally consumption data was observed every minute, and it has been further aggregated to hourly resolution for analysis.

[2]We use superscripts to index over consumers, and subscripts to index over other variables such as time.

with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$, $(z)_+ \equiv \max(z, 0)$ and $(z)_- \equiv \min(z, 0)$ for an arbitrary real value z, and x_0 is a non-thermal base load. For $T_t < T_0$, where T_0 is the thermostat setpoint, the premise will use the heating appliance (resulting in a negative response a_- of consumption with temperature), whereas for $T_t > T_0$ the premise will use the AC (with a positive response rate to temperature a_+). We estimate this model as in [2]. Under this model, the expected value of the thermal response rate for the given user, and at a specified temperature level T, is computed as $a(T) = a_-$ for $T < T_0$ and $a(T) = a_+$ for $T \geq T_0$.

The thermal regimes model. For a given consumer, we apply the model [3] to identify windows in the usage time series where either HVAC may be used, or no HVAC is in use. The model identifies J regimes (or states) of consumption, whereby each regime j, $j = 1, ..., J$, is characterized by parameters (b_j, a_j, σ_j), corresponding to the base load, the thermal response (in kW/°F), and the consumption volatility in that regime. Model parameters are estimated from the data $\{x_t, T_t\}_{t=1}^{\mathcal{T}}$ via a maximum likelihood procedure [3]. The model has two main components: *i)* a state-specific consumption response to temperature, and *ii)* a decision process that governs how the user switches between regimes.

Regime-specific response. If at time t the user is in regime j, consumption readings x_t are generated according to:

$$x_t = b_j + a_j(T_t - T_0) + \epsilon_j, \tag{2}$$

where $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ is a state-specific consumption volatility that is normally-distributed. We may interpret the case $a_j < 0$ as heating (the lower the temperature, the more energy is being used); $a_j > 0$ as cooling (the higher the temperature, the more energy is being used), and $a_j \approx 0$ as no HVAC usage.

Switching between regimes. We let $s_t = j$ denote that the given consumer is in regime j at time t. In that state, given a level of temperature T_t, the user chooses the next regime l, $l = 1, ..., J$, that maximizes a linear utility function depending on both the outside temperature and the current state:

$$U_j(T_t | s_t = l) = c_{jl} + d_{jl} T_t + \eta_{jl}, \tag{3}$$

which is a widely-used specification in random choice modeling [15]. Here, c_{jl} and d_{jl} are, in turn, the base utility and the rate of change with temperature from transitioning from regime j to regime l. Adopting the common assumption that the random variable η_{jl} follows an Extreme-Value distribution, the probability from transitioning to regime l when in regime j has a closed-form solution [15]:

$$P(s_{t+1} = l | s_t = j, T_t) = \frac{\exp(c_{jl} + d_{jl} T_t)}{\sum_m \exp(c_{jm} + d_{jm} T_t)}. \tag{4}$$

Above we made use of the Markov assumption that the state of the user at time $t+1$ only depends on its state at the current time step t, but not on all past history. Note that, while the thermal regimes model views consumption from the perspective of consumer decisions, the same mathematical formulation can be interpreted from the perspective of modeling the probability that a load control device will switch states given external stimuli (such as temperature).

The succession of states over time describes the user's usage behavior: $\{a_t\}_{t=1}^{\mathcal{T}} \equiv \{a_{s_t}\}_{t=1}^{\mathcal{T}}$ may suggest whether HVAC is being used at time t and should be compared with metrics derived from the separately-monitored HVAC h_t, whereas $\{\sigma_t\}_{t=1}^{\mathcal{T}} \equiv \{\sigma_{s_t}\}_{t=1}^{\mathcal{T}}$ corresponds to intentional consumption and should be contrasted with the occupant-driven usage o_t. Using this model, we may compute the expected value of the thermal response rate for the given user at a specified temperature level T, $a(T) = \sum_j P(s = j|T) a_j$.

The "true" thermal response. Using separately-monitored HVAC data, we compute the "true" linear response rate $\hat{a}(T)$ of a given consumer for a specified temperature level T using a locally-weighted regression:

$$h_t = \bar{h} + \hat{a}(T_t)(T_t - T_0) w_t + \nu_t, \tag{5}$$

with $\bar{h}$ the non-temperature-dependent HVAC usage, and ν_t an error term. The weights take a low-order polynomial form, $w_t \equiv 1 - \left| \frac{\sum_{t'} (T_t - T_{t'})}{d(T_t)} \right|^3$ (the *LOESS* model [16]), where only the observations $T_{t'}$ close in value to T_t, in the sense that $|T_{t'} - T_t| \leq d(T_t)$, contribute to the weighting. Here $d(T_t)$ is computed by the estimation algorithm in [16], and depends on the density of observations around a temperature level T_t.

III. Deciding enrollment for Demand-Response

A. Ranking consumers by flexibility potential

We define a ranking $\mathbf{r}$ of $n \leq N$ consumers as an ordered set of indices (i_1, ..., i_n), so that $i_k \neq i_{k'} \forall k \neq k'$ (i.e., we assume rankings are unique).

Problem formulation. If one had separately-monitored HVAC data h_t, one could rank the consumers $i = 1, ..., n$ according to their actual temperature response $\hat{a}^i$:

$$\hat{\mathbf{r}} \equiv (\hat{i}_1, \hat{i}_2, ..., \hat{i}_n) : \hat{a}^{\hat{i}_1} > \hat{a}^{\hat{i}_2} > ... > \hat{a}^{\hat{i}_n}.$$

However, only whole-home consumption readings x_t are typically available, and a statistical model will be used to estimate the potential flexibility a^i, which leads to a different ranking:

$$\mathbf{r} \equiv (i_1, i_2, ..., i_n) : a^{i_1} > a^{i_2} > ... > a^{i_n}.$$

The question we address here is how to compare $\hat{\mathbf{r}}$ and $\mathbf{r}$. For example, in the right panel in Table I, the top $n = 10$ consumers have been selected first by $\hat{a}(T)$, then by $a(T)$ for $T = 80$°F. The response values for all consumers and the relative positions of some of them, e.g., for Danielle, Arthur, and Frank, are different for a and $\hat{a}$. A distinction exists between programs that reward consumers for being part of a desirable group (e.g., the top 10% by some metric), for which the same incentive is being offered to each user in the group, and programs for which the relative performance within a peer group is relevant (e.g., incentives proportional to the ranking within the group). We discuss these two cases below.

Ranking by cumulative thermal response. For a given consumer, and at a given temperature T, we compute the average ground-truth thermal response $\hat{a}(T)$ and the model-estimated response $a(T)$ for a narrow range around T (e.g., $[T - 1°\text{F}, T + 1°\text{F}]$). For simplicity, we assume that the n users are independent. We next order the consumers either by the model-estimated response, obtaining the ranked list $\mathbf{r} = (i_1, ..., i_n)$, or by their ground truth response, obtaining the ranked list $\hat{\mathbf{r}} = (\hat{i}_1, ..., \hat{i}_n)$. Using these ranked lists, we

TABLE I
RANKED LISTS AND MODEL-ESTIMATED RESPONSE a AND "TRUE" RESPONSE $\hat{a}$ FOR $T = 80^\circ$ F.

	Ranking by a	$a(80^\circ F)$	Ranking by $\hat{a}$	$\hat{a}(80^\circ F)$
5	Arthur	0.011	Danielle	0.137
3	Bill	0.014	Helen	0.106
4	Carla	0.012	Bill	0.086
1	Danielle	0.017	Carla	0.085
7	Edd	0.007	Arthur	0.080
6	Frank	0.009	Frank	0.062
9	George	0.015	Edd	0.054
2	Helen	0.005	James	0.007
10	Irene	0.006	George	0.006
8	James	0.002	Irene	0.000

compute the cumulative sum of the total *ground truth* response for the top n consumers, when ordered either by the model-induced ranking, or by the true ranking:

$$S_n = \sum_{l=1}^{n} \hat{a}^{i_l} \text{ and } \hat{S}_n = \sum_{l=1}^{n} \hat{a}^{\hat{i}_l}. \tag{6}$$

for different values of $n = 1, ..., N$. Then, $\hat{S}_n$ represents the total "true" energy reduction obtained by enrolling the top n consumers when ground truth data is available for this purpose, whereas S_n represents the "true" response when ground truth data is not available, and the consumers are ranked by the response values estimated using the model. This setup internalizes the simple fact that, regardless of whether ground truth data is available or not, the actual flexibility obtained by enrolling consumers in any order is still given by their true response. We would like to empirically evaluate the mismatch $\Delta_n \equiv \hat{S}_n - S_n$ between the true cumulative responses computed by the two methods, and the corresponding relative absolute error $E_n = \frac{|\Delta_n|}{\hat{S}_n}$.

Comparing ranking structures. Consider two ranked lists of n consumers, $\mathbf{r}$ and $\hat{\mathbf{r}}$. The pair of consumers (l, l') is said to be *concordant* if *i)* $\hat{a}^l < \hat{a}^{l'}$ and $a^l < a^{l'}$ or *ii)* $\hat{a}^l > \hat{a}^{l'}$ and $a^l > a^{l'}$; otherwise the pairs are called *discordant*. A popular metric for assessing the similarity of rankings defined this way is the τ statistic, introduced by Kendall [17] as:

$$\tau = \frac{C - D}{C + D}, \tag{7}$$

where $C \equiv \sum_{l,l'=1}^{L} \mathbb{I}\{\text{pair } (l, l') \text{ is concordant}\}$ and $D \equiv \sum_{l,l'=1}^{L} \mathbb{1}\{\text{pair } (l, l') \text{ is discordant}\}$. Rankings $\mathbf{r}$ and $\hat{\mathbf{r}}$ that are in perfect agreement yield $\tau = 1$; for perfect disagreement $\tau = -1$. If there is no relation between the rankings, $\tau = 0$.

The τ statistic admits a non-parametric hypothesis test, which for large n (> 10) may be approximated by a z-test given that τ will follow a zero-mean normal distribution with variance $\sqrt{\frac{2(2n+5)}{9n(n-1)}}$ [17]. This may be then used to compute p-values of the obtained estimations of τ under the null hypothesis $\mathcal{H}_0 : \tau = 0$. We may reject $\mathcal{H}_0$ that the rankings $\hat{\mathbf{r}}$ and $\mathbf{r}$ are independent for small values of p.

While comparing rankings using Kendall's τ (or similar statistics) is easy and intuitive, there are several issues with its definition that limit the analysis for our application:

1) All user pairs (l, l') are given equal weight, even if their respective response rates are widely different.

2) All user pairs (l, l') are considered statistically independent. Yet this is not true, since if consumers l and l' are "similar" (that is, have similar thermal consumption patterns), they will tend to be ranked in a similar way, and, as such, one swap between the ranks of (l, l') will contain information about a potential swap of l' with another consumer l''.

3) Rankings according to thermal response should not be completely independent across similar temperature ranges. For example, rankings at $T = 70^\circ$ F and $T = 71^\circ$ F should be relatively close, perhaps more so than at $T = 90^\circ$ F.

A modified rank distance. To address the above issues, we adapt the approach in [18] to compute the probability of observing a ranking $\mathbf{r}$ when given a "true" ranking $\hat{\mathbf{r}}$, and where there exist correlations between rankings. The rank distance d_{rank} introduced there *i)* penalizes swaps between users that are very different in their thermal response more than swaps between users that have similar thermal response, and *ii)* penalizes swaps between pairs of users conditional on swaps among the other user pairs.

We discretize temperature values into Q bins, such that the density of observations is similar across bins. Following [18], for each bin q, $q = 1, ..., Q$, we define a $3 \times n$ matrix $A_q \equiv (a^i(T_{q,\text{left}}), a^i(T_{q,\text{middle}}), a^i(T_{q,\text{right}}))_{i=1}^n$, where the temperature values correspond to the left, middle, and right points of the bin q (rounded to the closest integer value for simplicity). We denote by μ_q the column mean vector of A_q ($\mu_q \in \mathbb{R}^n$). We also define $y_q \equiv (\hat{a}^i(T_{q,\text{middle}}))_{i=1}^n$, the average ground truth response for a given temperature interval q, and order y_q, μ_q, and the columns of A_q by decreasing values of y_q. Then the distance between the rankings induced by A_q and y_q is:

$$d^2_{\text{rank}}(y_q, A_q) \equiv \min_{\boldsymbol{\lambda} > \mathbf{0}} (\lambda - \mu_\Delta(y_q)) \Sigma_\Delta^{-1}(A_q) (\lambda - \mu_\Delta(y_q))^T \tag{8}$$

where $\mu_\Delta(y_q)$ is the vector obtained by taking differences between adjacent values in μ_q after ordering its rows by y_q, and $\Sigma_\Delta(A_q)$ is the covariance matrix of the differences in response across the columns of A_q. Here $\boldsymbol{\lambda}$ is a vector in $\mathbb{R}^{N-1}$ that has maximum probability under a distribution defined by A_q and which also preserves the ranking of y_q. There is a direct correspondence [18] between d^2_{rank} and Hotelling's T^2, which is inversely proportional to the probability of observing $\boldsymbol{\lambda}$ when the true mean is y_q. The distance between rankings defined in this way admits a non-parametric hypothesis test $\mathcal{H}'_0 : d_{\text{rank}} = 0$ (the rankings are the same). In [18], a bootstrapping-based procedure is outlined for computing the p-value of such a test. If the computed p is smaller than a threshold, we may reject $\mathcal{H}'_0$.

Example. The lists in Table I are similar in the relative rankings, but there certainly are differences. For the ranked lists in the right panel, Kendall's τ is computed as $\tau = 0.46$ ($p = 0.07$), whereas $d_{\text{rank}} = 2.14$ ($P(d_{\text{rank}} = 0) = 1$). Judging by τ ($\tau < 0.5$, $p > 0.05$), the two rankings are different. However the question we ask is how similar *the processes* are that generated these rankings–and for this purpose d_{rank} is better suited. On this example, d_{rank} indicates a very good similarity between the processes that generated a and $\hat{a}$.

B. Detecting occupancy and HVAC use

For DR targeting, it is useful to gauge the likelihood that a consumer will adopt a request for effort. Depending on how this request is made (e.g., automatic or requiring human action), the fact that the consumer is at home or not is likely to make a significant difference on compliance. E.g., if the request is to turn the HVAC setpoint up manually during certain hours, the consumer cannot comply if he is not at home during that time. If the DR event automatically adjusts the setting of a smart thermostat, this may interfere with the comfort level of a consumer who is at home at that time.

As such, we analyze the use of the thermal regimes model as a detector of both HVAC usage and of *occupancy*, here understood as periods of time when significant occupant-induced consumption is observed. As ground-truth we use the separately-monitored, occupant-induced consumption activity $o_t^i, i = 1, ..., n, t = 1, ..., \mathcal{T}^i$ (`user` from Table II). We would like to compare the detection decisions made using thermal model results with those made when separately-monitored h_t^i and o_t^i data is available. These decisions are of the form:

h_t^i / o_t^i	**Low**	**High**
Low	HVAC off, Inactive	HVAC off, Active
High	HVAC on, Inactive	HVAC on, Active

In practice, simple rules are often used for these decisions. We define such rules for decisions based either on estimated parameters identified through the thermal model (TR), or on ground-truth HVAC or occupant-driven usage (GT):

$$\delta_{\text{TR}}(a_t) = \begin{cases} \text{High} & \frac{a_t}{\min_j a_j} \geq \lambda \\ \text{Low} & \text{otherwise} \end{cases}; \delta_{\text{GT}}(\hat{a}_t) = \begin{cases} \text{High} & \hat{a}_t \geq q_{\hat{a},\gamma} \\ \text{Low} & \text{otherwise} \end{cases}$$

The same rules apply for σ_t. Here, γ and λ are decision parameters ("knobs") that an operator may tweak in order to adjust how large an observed quantity should be in order for it to be considered significant. E.g., the higher γ is, the fewer times the utility will consider that the observed separately-monitored consumption corresponds to significant HVAC use or user activity. Thus, for every consumer i we study how often the thermal regimes model arrives at the same decision (e.g., "HVAC on, Active") as when actual ground truth data is available, i.e., compute the detection accuracy

$$acc^i = \frac{1}{\mathcal{T}^i} \sum_{t=1}^{\mathcal{T}^i} \mathbb{I}(d_{TR}(a_t^i) = d_{GT}(\hat{a}_t^i)). \quad (9)$$

IV. Experimental setup

A. The Pecan St. data

The Pecan St experiment [19] is a large smart grid research deployment in Austin, TX, where 500 homes are equipped with sensors that collect minute-level electricity use data, both at the whole-home level, and for 8 to 23 individually-monitored appliance circuits. We further aggregate these into the general categories described in Table II, and to an hourly level. Also presented are the shares of consumers in the final dataset that had different types of end-uses monitored. For our analysis, we used whole-home and individually-monitored consumption time series from 336 consumers who had HVAC appliances.

TABLE II
Categories of end-use for the Pecan St. experiment [19].

End use	**Category**	**Appliances in this category**	**Monitored**
`H`	Thermal	Space Heater	85%
`C`	Thermal	Air, Air Window Unit, House Fan	92%
`user`	Occupant	Bathroom, Bedroom, Clothes Washer, Dryer, Garage, Kitchen, Lights Plugs, Outside Lights, Kitchen, Microwave, Dining Room, Oven, Security	98%

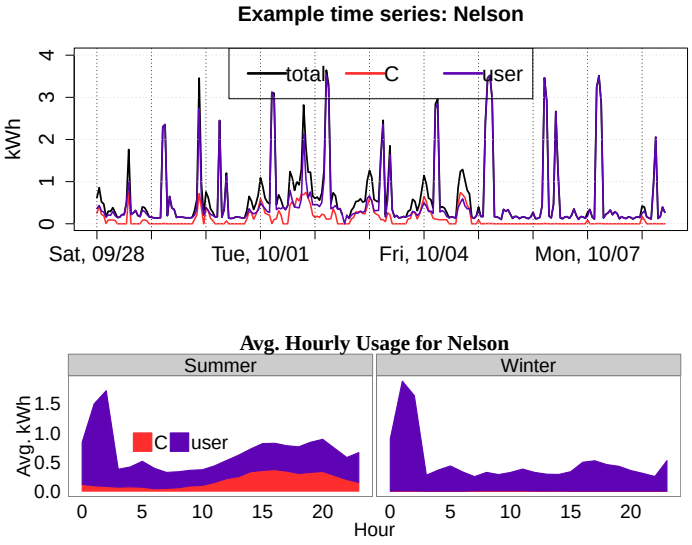

Fig. 2. End-use consumption data example for one example user, "Nelson". *Top:* A week of hourly consumption data across separately-monitored end-uses; *Bottom:* Average seasonal consumption for summer (left) and winter (right). The HVAC and user-induced consumption components are plotted.

In Fig. 2, the categories defined in Table II are plotted for one example consumer, "Nelson". The left panel presents consumption for a sample one-week window; the right panel shows a year's worth of data as averages over time-of-day and seasons. Note the increased AC use in the summer (as opposed to the winter), and the absence of heating, even in the winter (Austin, TX winters are mild and summers are hot). Intentional, occupant-driven consumption behavior (`user`) is consistent throughout the day across both summer and winter.

For each consumer, we have obtained weather data for Austin, TX, using an on-line API (www.wunderground.com). We aggregated the 5-minute readings to an hourly resolution.

B. Individual thermal consumption models: example

We learn the model described in Section II for the final dataset of $N = 336$ consumers who had a thermal appliance monitored. Estimation is done as described in detail in [3], following a cross-validation procedure that finds the smallest model (in terms of the number k of thermal regimes) that explains at least $R = 85\%$ of the out-of-sample variance.

In Fig. 3, we illustrate model results, for comparison, for three example users, "Jose", "Peter", and "Nelson" (from left to right). The two darker-hue lines in the figure show the temperature profiles of the thermal response, $a(T)$, computed from

whole-home data when using either the thermal regimes model (TR) or the linear breakpoint response (BP) models, whereas the lighter color line corresponds to the ground truth profile (GT) computed from separately-monitored HVAC appliance data, $\hat{a}(T)$. For each of the example consumers, the mismatch between the model estimate and the true response is evident, although the TR curve will tend to match the ground truth more closely. However we are less interested in the absolute value of the response, as in the use of these estimates for ranking, as illustrated in Section V.

V. Experimental results

A. Ranking users for program enrollment

By sorting consumers in the decreasing order of the absolute temperature response rate $a(T)$, at a given temperature level T, the DR program identifies those households who can provide the most demand-side flexibility (energy savings) by changing their thermostat setpoint (either up or down) by 1°F. We highlight the conditions under which the model is most likely able to yield good rankings of consumers (when compared with the rankings based on ground truth response), even if the numerical estimates of the response rates may be different under the model than the true ones.

Comparing rankings by cumulative thermal response. The first question we ask is how similar the true cumulative response of the top n consumers is when sorting according to the estimated response rates $a^i(T), i = 1, ..., N = 336$, at a given temperature level T, to the situation when sorting according to the true response $\hat{a}^i(T)$. As such, we compute the cumulative response $S_n(T)$ (see (6)), $n = 1, ..., N$, for levels of $T = 40, ..., 120$°F. The light-colored curve in Fig. 4 illustrates the cumulative thermal response $\hat{S}_n$ as a function of the number n of top consumers, selected by their true temperature response for three representative temperature levels, $T_1 = 40$°F, $T_2 = 80$°F, and $T_3 = 110$°F (from left to right). At the latter two temperature levels, the thermal appliance is likely to be in use for most consumers. In each panel, we overlay the cumulative thermal response S_n obtained when ranking consumers by model estimates obtained either by the thermal regimes model (TR, dark, solid line), or by the linear breakpoint model (BP, dashed dark line). The light-colored, dotted line indicates the "true" cumulative response obtained in the case when consumers are selected at random, without a prior ranking operation. The effects of the ranking $\mathbf{r}$ achieved using a are indeed quite close to those obtained using $\hat{a}$. Moreover, the cumulative ranking strategy is clearly superior to the random selection strategy, with gains over the random case in excess of 100% in the estimated averted consumption for each 1°F change in the thermostat setpoint for medium values of n.

The relative error $E_n(T)$ is plotted in Fig. 5 as a function of the temperature level T and the size of the list of interest n, when using, in turn, the thermal regimes model (TR, top panel), and the linear breakpoint model (BP, lower panel). The figures present heatmaps of $E_n(T)$ where dark hues are regions of low relative error and light hues regions of high relative error. Obviously, for both models, the more consumers are selected, the lower the relative error (since in the end all $N = 336$ are selected, regardless of the selection order). However, noteworthy patterns emerge, that allow to contrast the ranking performance of the two models. For intermediate temperatures $T \in [65$°F, 85°F], the relative errors obtained using the thermal regimes model are generally under 30% for most values of $n = 5, ..., N$, whereas they stay relatively constant across temperatures for the linear breakpoint model. This temperature range is generally associated with moderate HVAC usage, which suggests that, on average, the thermal regimes model captures those situations when the thermal appliances are not in heavy use (it thus acts as a good detector of HVAC usage), whereas the breakpoint model is insensitive to this. For temperatures $T \gtrsim 85$°F and a medium-size enrollment group $n \in [20, 70]$, E_n for the thermal regimes model is typically within a range of $5\% - 10\%$, whereas for the same-size enrollment groups but low temperatures ($T < 65$°F) the obtained relative errors are $30\% - 50\%$. In contrast, the error levels for the breakpoint model are consistently above $65\% - 70\%$ regardless of the temperature level. This indicates that the thermal regimes model is better able to estimate actual temperature response for both cooling and heating regimes than the breakpoint model, while itself being better at identifying cooling rather than heating. This is an expected behavior, as cooling is a sizeable component of the total energy budget of a typical residential consumer in a hot area like Austin, TX.

Comparing ranking correlations. As argued in Section III, certain applications may require understanding the actual structure of the ranked $n-$list obtained, not just the cumulative kWh savings from the top n consumers. E.g., this could be the case of a program that financially rewards individuals for their performance compared with that of "similar" consumers. To study how similar the lists $\mathbf{r}$ and $\hat{\mathbf{r}}$ actually are, we use the rank correlation measure d_{rank} proposed in Section III. In particular, we analyzed the quality of the match between the ranked lists obtained by sorting consumers, in turn, by their true temperature response $\hat{a}$ and by model response a (obtained using either the thermal regimes or the linear breakpoint models, see Fig. 6).

We computed the $d_{\text{rank}}(r, \hat{r})$ distance measure as described in Section III across close temperature levels and for different sizes n of the enrollment list. In Fig. 6 we illustrate the normalized $d_{\text{rank}}(r, \hat{r}; T, n)$ (across the ns) for given levels of the temperature T, for both the thermal regimes model (left panel), and the linear breakpoint model (middle panel). Note that for d_{rank}, the lower the value, the closer the two rankings $\mathbf{r}$ and $\hat{\mathbf{r}}$. For the thermal regimes model, there is a good match between the rankings ($d_{\text{rank}} \lesssim 0.1$) for higher temperature ranges ($T = 75°F...110$°F), and for relatively small list sizes $n \lesssim 25$. As expected intuitively, the normalized distance increases with n, as a structural match between the rankings is harder to achieve when many consumers are involved. However, for the linear breakpoint model the match is much worse, with normalized d_{rank} values consistently over 0.7. The best results for the breakpoint model are achieved around $T = 70$°F.

Next, focusing on the thermal regimes model, we computed the p-values of hypothesis $\mathcal{H}_0 : d_{\text{rank}}(r, \hat{r}; T, n) = 0$. Intuitively, if many samples were to be taken of thermal

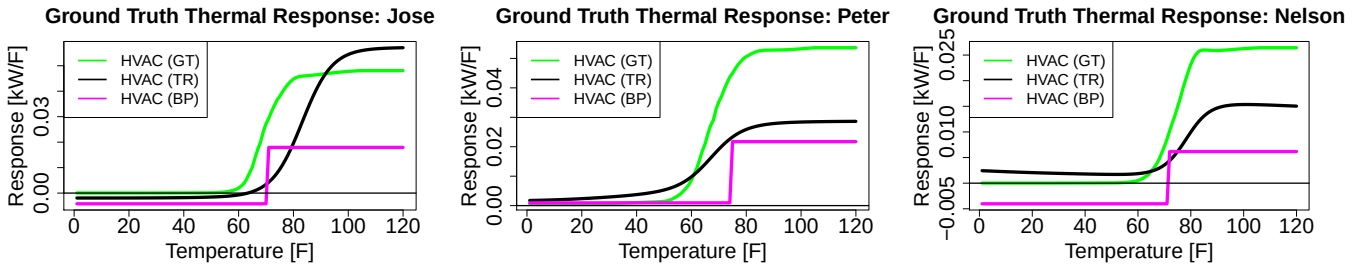

Fig. 3. Ground truth (GT) and estimated response profiles – either by the thermal regimes (TR) model or by the linear breakpoint (BP) model – for three example users, "Jose", "Peter", and "Nelson". The TR curve will generally match the ground truth better than the BR one.

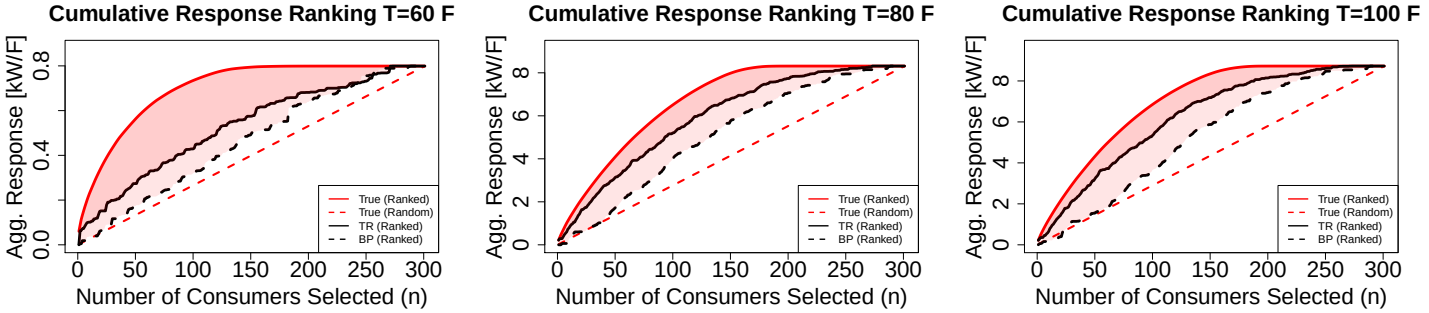

Fig. 4. Cumulative thermal response (energy savings) obtained by sorting consumers according to thermal regimes model-estimated thermal response for three temperature levels (60°F, 80°F, and 110°F, from left to right). The "true" estimated averted saved energy (cumulative response) obtained by ranking consumers according either to the ground truth temperature response (solid, light), to the thermal regimes estimates (TR, solid dark), or to the breakpoint model estimates (BP, dashed dark) are depicted. The dotted light line shows the estimated averted consumption under a random selection strategy.

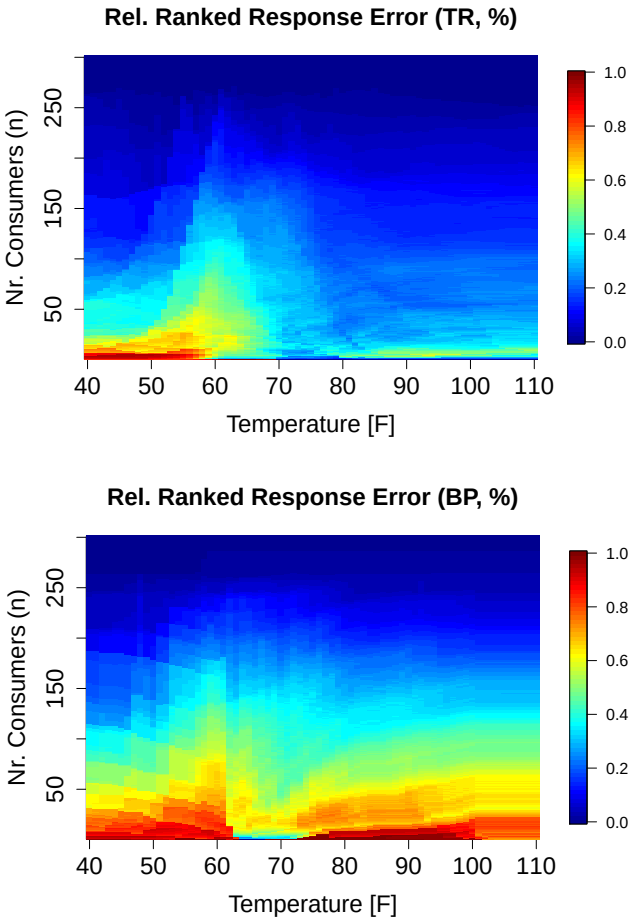

Fig. 5. Relative error (on a $0-1$ scale) $E_n(T)$ between true savings obtained by sorting consumers according to the thermal regimes model and the savings obtained when ground truth data was available.

TABLE III
OCCUPANCY (ACTIVE OR INACTIVE) AND HVAC DETECTION (ON OR OFF) STATISTICS FOR NELSON FOR DECISION THRESHOLD PARAMETERS $\theta = 2, \gamma = 0.5$. *Left*: THERMAL REGIMES MODEL; *Middle*: GROUND TRUTH DATA; *Right*: AGREEMENT BETWEEN DETECTION DECISIONS.

	Off	On
Inact.	0.36	0
Act.	0	0.64

	Off	On
Inact.	0.31	0.24
Act.	0.19	0.26

	Off	On
Inact.	0.80	0
Act.	0	0.97

response estimates for temperature values around a given level T (obtained by applying the thermal regime model for these different temperature levels), $P(d_{\text{rank}} = 0)$ indicates the probability that the processes that have generated the two rankings $\mathbf{r}$ and $\hat{\mathbf{r}}$ are functionally identical. We illustrate this in the middle panel in Fig. 6 using the same grayscale heatmap, where larger (darker) values indicate better rankings. As before, we have discretized the $[0, 1]$ interval into three bins, with Low : $p \in [0, 0.90]$, Medium : $p \in (0.90, 0.95]$, High : $p \in (0.95, 1.00]$, corresponding to the probability of rejecting $\mathcal{H}_0$ when the hypothesis is actually true. This profile suggests good structural matches between the ground-truth and model-estimated rankings for high temperature levels ($T > 85$°F) for low to medium enrollment list sizes ($K \lesssim 40$). For low temperatures ($T \lesssim 80$°F), the structural ranking match is consistently high only for small enrollment list sizes ($n \lesssim 20$), and sparsely for larger n, indicating a degree of instability in the estimation of thermal response for low T.

B. Detecting occupancy

In Table III, we illustrate the detection statistics for Nelson for selected decision thresholds $\gamma = 0.5$ and $\theta = 2$. The left and middle panels show, in turn, the share of hours across two years of data when Nelson is estimated to be in either of the

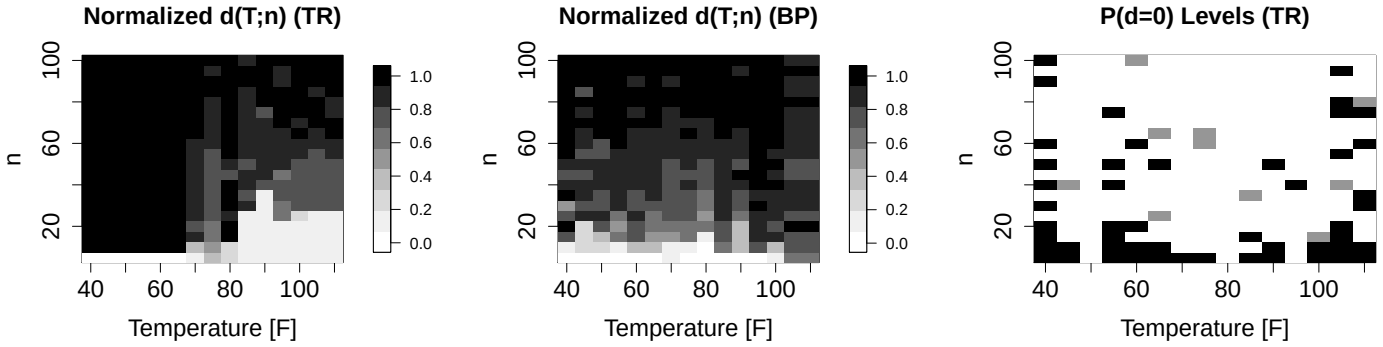

Fig. 6. Ranking correlation measures as a function of temperature T and list size n. *Left:* and *Right:* Normalized d_{rank} for each value of T (lower is better) for the thermal regimes (TR) and linear breakpoint (BP) models; the values in this panel are in the $[0, 1]$ interval. *Right:* $P(d_{\text{rank}}; T, n) = 0$ profile (discretized in Low/Medium/High bins, depicted in white/gray/black) for the thermal regimes model.

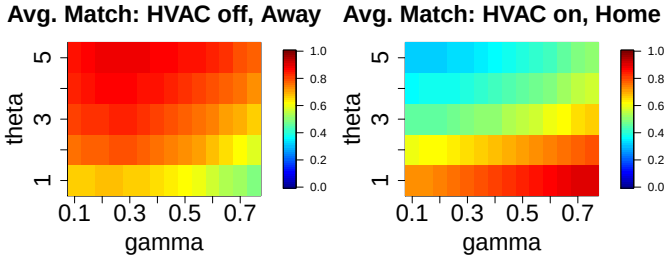

Fig. 7. The dependence on decision parameters θ and γ of the average agreement (AA) between decisions made with the thermal mode and the ground-truth data. Left: "HVAC off, Inactive"; *Right*: "HVAC on, Active".

four occupancy regimes ("HVAC off, Inactive", "HVAC off, Active", "HVAC on, Inactive", "HVAC on, Active") by either the thermal-regimes model (d_{TR}, left) or the ground-truth data (d_{GT}, middle). The right panel shows the match between the detection decisions d_{GT} and d_{TR}. Two observations are in order for Nelson: 1) the thermal regimes does poorly in identifying the "HVAC off, Active" and "HVAC on, Inactive" occupancy states, and 2) it performs fairly well in identifying the "HVAC off, Inactive" and "HVAC on, Active" states. Note that these numbers will change with θ and γ, i.e., with what constitutes a high-enough HVAC or occupant activity to be marked as such by the simple decision rules.

Fig. 7 illustrates the empirical dependence on the decision parameters θ and γ of the average agreement ($AA(\theta, \gamma) \equiv \frac{1}{N} \sum_i acc^i(\theta, \gamma)$) between two types of detections that are very relevant in practice, "HVAC off, Inactive" and "HVAC on, Active". Certainly, there is little use of sending a request for effort (e.g., a manual thermostat setpoint change) to a consumer who both might be away and not using HVAC at the same time; however, if the consumer is at home (inferred from the proxy consumption activity) and using HVAC, sending the request is advisable. Interestingly, the agreement between d_{GT} and d_{TR} differs in its dependence on θ and γ for the two states. For the "HVAC off, Inactive" state, agreement increases with higher θ but for intermediate values of γ, which is surprising, since one would expect agreement to increase with γ. Conversely, for the "HVAC on, Active" state, the agreement between the model and the ground truth-based decisions increases with γ, but decreases with θ. These sensitivity maps suggest that there is a trade-off between the detection performance of the different states of interest, and that appropriate values for the detection parameter θ should be chosen to correspond to agreeable ranges of the tolerance γ.

VI. Discussion

In this paper we develop a methodological framework to assess the performance of statistical models designed to disaggregate individual thermal energy use from whole-home smart meter data. Our motivation is when the end-goal of the model is designing strategies for enrolling consumers into thermal DSM programs. We frame the problem as one of comparing ranked lists of estimated and "ground truth" individual thermal responses at different temperature levels. We illustrate this methodology through an experimental evaluation of two prior consumption models and their use for ranking and selecting groups of consumers for thermal demand-response.

The evidence presented here suggests that a simple model such as [3], if carefully used, may appropriately guide demand-response programs, removing the need for both acquiring appliance-level data (which may be expensive and intrusive to obtain) and for using potentially more complex models (which are typically hard to interpret and more computationally-intensive). It performs better in obtaining good rankings than an even simpler linear breakpoint model [9], [2]. This is not surprising, as the thermal regimes model offers more flexibility in identifying temperature-dependent thermal response rates. To be practical and useful for thermal DSM program enrollment, a model needs not be perfect, but generate similar rankings as when using ground truth data.

The analysis also highlights the limitations of the thermal models employed here, mainly their simplicity in assuming the decision processes of HVAC usage and user-generated activity. Because the thermal regimes model is based on states characterized *jointly* by temperature response levels and occupancy-generated usage, it effectively couples the types of states of the two end-uses that can be detected, which are capped by the number of regimes assumed in the model. Moreover, we have limited our analysis to the expected value of the thermal response, and ignored the correlations between the consumption patterns of users. This is especially important when evaluating the error profile of the estimated aggregate

thermal flexibility. We address that in a separate work on the covariance structure of consumer groups.

While the size of the data sample used here is relatively small, it is typical for a small urban neighborhood serviced by a distribution substation. Within a larger customer base, customer enrollment for demand-response should be performed separately and independently for each local consumer population (e.g., served by the same substation and experiencing similar weather). Scaling the evaluation method presented here to a large customer population entails assessing separately the quality of each of the enrollment groups selected from local populations. This hierarchical design of DSM programs is a rich topic for research to which we anticipate to contribute in future work.

REFERENCES

[1] S. Barker, S. Kalra, D. Irwin, and P. Shenoy, "Nilm redux: The case for emphasizing applications over accuracy," June 2014.

[2] V. M. Muggeo, "Estimating regression models with unknown break-points.," *Statistics in Medicine*, vol. 22, pp. 3055–3071, 2003.

[3] A. Albert and R. Rajagopal, "Thermal profiling of residential energy use," *Power Systems, IEEE Transactions on*, vol. 30, no. 2, 2015.

[4] J. Eto, J. Nelson-Hoffmana, E. Parker, C. Bernier, P. Young, D. Sheehan, J. Kueck, and B. Kirby, "Demand response spinning reserve demonstration?phase 2 findings from the summer of 2008," *Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA*, 2009.

[5] J. L. Mathieu, M. E. Dyson, and D. S. Callaway, "Resource and revenue potential of california residential load participation in ancillary services," *Energy Policy*, vol. 80, pp. 76–87, 2015.

[6] R. S. J. Tol, S. Petrick, and K. Rehdanz, "The impact of temperature changes on residential energy use," Working Paper Series 4412, Department of Economics, University of Sussex, Nov. 2012.

[7] O. G. Santin, L. Itard, and H. Visscher, "The effect of occupancy and building characteristics on energy use for space and water heating in dutch residential stock," *Energy and Buildings*, vol. 41, no. 11, 2009.

[8] M. E. Dyson, "Using smart meter data to estimate demand response potential, with application to solar energy integration," *Energy Policy*, vol. 73, pp. 607–619, 2014.

[9] B. J. Birt, G. R. Newsham, I. Beausoleil-Morrison, M. M. Armstrong, N. Saldanha, and I. H. Rowlands, "Disaggregating categories of electrical energy end-use from whole-house hourly data," *Energy and Buildings*, vol. 50, no. 0, pp. 93 – 102, 2012.

[10] D. Huang, M. Thottan, and F. Feather, "Designing customized energy services based on disaggregation of heating usage," in *Innovative Smart Grid Technologies (ISGT), 2013 IEEE PES*, pp. 1–6, 2013.

[11] E. C. Kara, M. D. Tabone, J. S. MacDonald, D. S. Callaway, and S. Kiliccote, "Quantifying flexibility of residential thermostatically controlled loads for demand response: a data-driven approach," *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings*, pp. 140–147, 2014.

[12] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *Power Systems, IEEE Transactions on*, vol. 28, pp. 4019–4030, Nov 2013.

[13] C. K. Carmel, G. Shrimali, and A. Albert, "Disaggregation: the holy grail of energy efficiency?," *Energy Policy*, 2013.

[14] J. Z. Kolter and T. Jaakkola, "Approximate inference in additive factorial hmms with application to energy disaggregation," *Journal of Machine Learning Research - Proceedings Track*, vol. 22, pp. 1472–1482, 2012.

[15] M. Ben-Akiva and S. R. Lerman, *Discrete choice analysis*. MIT Press, 1985.

[16] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, pp. 596–610, 1988.

[17] H. Abdi, "The kendall rank correlation coefficient," in *Encyclopedia of Measurement and Statistics*, pp. 508–510, Sage, 2007.

[18] B. Carterette, "On rank correlation and the distance between rankings," in *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, (New York, NY, USA), pp. 436–443, ACM, 2009.

[19] B. McCracken, M. Crosby, C. Holcomb, S. Russo, and C. Smithson, "Data-driven insights from the nations deepest ever research on customer energy use," technical report, Pecan Street Research Institute, 2013.