"This is an interesting application paper and I enjoyed reading it."
— Meta Reviewer

Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data *

Tri Kurniawan Wijaya [†‡] Tanuja Ganu [§] Dipanjan Chakraborty[§] Karl Aberer[†]
Deva P. Seetharam[§]

Abstract

Many electricity suppliers around the world are deploying smart meters to gather fine-grained spatiotemporal consumption data and to effectively manage the collective demand of their consumer base. In this paper, we introduce a structured framework and a discriminative index that can be used to segment the consumption data along multiple contextual dimensions such as locations, communities, seasons, weather patterns, holidays, etc. The generated segments can enable various higher-level applications such as usage-specific tariff structures, theft detection, consumer-specific demand response programs, etc. Our framework is also able to track consumers' behavioral changes, evaluate different temporal aggregations, and identify main characteristics which define a cluster.

1 Introduction

Many electricity suppliers around the world are deploying smart meters to gather fine-grained spatiotemporal consumption data [23]. These companies are interested in mining the collected data to extract deep insights such as *the set of consumers to be selected for winter peak load reduction, the set of consumers to be monitored for potential theft/anomaly, the set of consumers who can be targeted for energy efficiency programs, etc* [4, 16, 17]. These insights are necessary for multiple application sub-domains in the energy sector such as billing, energy audit, etc. For all such advanced applications, *consumer segmentation* has been viewed as one of key requirements [5, 10, 13, 15, 19].

However, segmenting consumers based on the smart meter data is challenging due to three reasons. First, the scale of smart meter data is humongous: high volume (data from millions of consumers) and high velocity (meters can report data at the rate of once every minutes to once every 30 minutes). Second, since electricity consumption is influenced by internal (family size, work hours, economic status, etc) and external contextual factors (weather, holiday, day of week, etc), the meter data must be correlated with heterogenous data sources (weather sites, survey results etc) that provide data at different time granularity and with varying data quality. Third, for meaningful grouping of consumers, the segmentation may need to be performed along disparate contextual dimensions.

To address these challenges, we propose a novel framework for consumer segmentation. The key contributions of this paper are:

- Design and implementation of a versatile framework for consumer segmentation that tries to jointly derive 'meaning' from consumption data, context data and user surveys. Previous works have primarily targeted a specific problem (e.g. setting tariff [11, 21], predicting consumer characteristics [3]) and do not consider this task in a holistic manner.
- Design of a temporal aggregation method that varies the level of aggregation based on application requirements and data quality.
- Design of a novel *clustering consistency index* to track the evolution of consumption behaviors (that helps spotting fraudulent activities such as thefts and tampering).
- Design of a novel *discriminative index* and survey mining approach to identify main consumer characteristics that can be used to classify consumers into well-demarcated clusters.

The rest of the paper is organized as follows. In Section 2, we give a brief overview of the related work. In Section 3, we explain our general purpose consumer segmentation framework. In Section 4, we discuss our clustering consistency index. In Section 5, we outline our method to extract knowledge from survey data to obtain main characteristics of a cluster. In Section 6, we describe the dataset and experimental results, and conclude in Section 7.

2 Related Work

Since the rollout of first batch of smart meters, smart meter data analytics has attracted immense interest. In particular, consumer segmentation has been considered to be crucial for enabling various smart grid applications. Moss et al. provided an overview of consumer segmentation in the electricity sector, especially for demand-side management [15]. Application-specific segmentations frameworks have been developed in [5, 6, 7, 10, 11, 19, 21], mainly for setting tariff or consumer classification, and predicting consumer characteristics [3].

In their work, Albert and Rajagopal also correlated consumers' consumption profile with demographics and appliance usages [3]. However, they did not attempt to iden-

*Supported by European Union's Seventh Framework Programme (FP7/2007-2013) 288322, WATTALYST.

[†]School of Computer and Communication Sciences, EPFL, Switzerland. {tri-kurniawan.wijaya,karl.aberer}@epfl.ch

[‡]The work is done during the author's internship at IBM Research, India.

[§]IBM Research, India. {tanuja.ganu,cdipanjan,dseetharam}@in.ibm.com

tify relevant characteristics for clustering consumers. Instead, they started with a predefined set of characteristics and determined whether those can be predicted from consumer's consumption profile. This approach is rather similar with [12] where the authors used the same dataset as ours to predict household demographics from consumption profiles. Regarding consumer behavior analysis, [18] presented a psychographic consumer segmentation based on how consumers feel, think, and act. However, their segmentation is based solely on survey data about consumers' behavior and attitude toward electricity and energy conservation, and did not involve processing of any consumption data.

Moreover, our *clustering consistency index* is related to Rand index [20]. However, we generalize it further to determine whether an individual is likely to change cluster over time, i.e., individual to cluster consistency measure.

Though, the earlier works have taken initial steps in deriving consumer segmentation based on smart meter data, they primarily target specific challenges/applications and do not present a 'general-purpose' framework. Moreover, none of the prior works have incorporated additional context-specific data for consumer segmentation. More importantly, as we will explain in the next section, different features and algorithms that are employed in the prior work, can be expressed as part of the building blocks in our framework.

3 Context-Specific Consumer Segmentation

In this work, we developed a context-specific 'general purpose' consumer segmentation framework which exhibits 5 design principles addressing bottom-up data federation challenges and top-down unifying solution requirements discussed earlier. This is a user-centric design which provides a certain freedom to the framework users to respecify the data, context, and feature spaces she is interested in based on the specific consumer segmentation task at hand. The framework enables the user to accomplish a number of consumer segmentation tasks such as segmenting the consumers based on the consumption magnitude, variability, or trend etc.

3.1 Design Principles We define the requirement specification based on 5 design principles which can be used individually or in combination by the framework user:

R1 (Customized Data Selection): to declare *a*) a period she is interested in, such as from June to August 2013 *b*) a subset of consumers that satisfy certain criteria (like average daily consumption greater than 5 kWh) *c*) a specific time of day she is interested in, such as afternoon peak hours 12:00 to 16:00

R2 (Customized Temporal Aggregation): to declare the time granularity of consumers' consumption profile used for segmentation, such as hourly, every 3 hours, daily, weekly, or monthly

R3 (Customized Context): to declare specific context such as summer, winter, weekend, January, Tuesday, temperature more than a certain threshold, etc.

R4 (Customized Features): to declare specific feature computation such as mean, standard deviation, coefficient of variation or median.

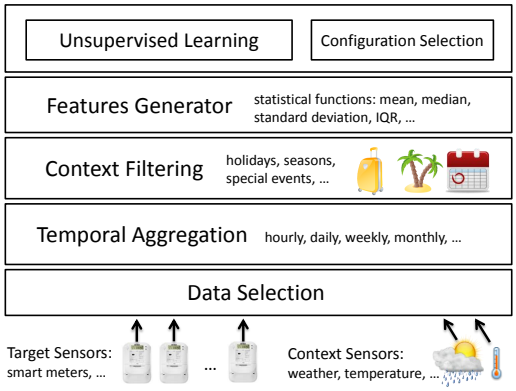


Figure 1: Architecture diagram of the framework.

R5 (Customized Algorithm) *a*) if the user has knowledge about a specific clustering algorithm to be used (from a predefined set) *b*) if supported by the algorithm, the user should be able to declare the number of consumer segments (or clusters) that she is interested in, or let the framework determine the best number of clusters (according to some cluster evaluation metrics).

3.2 General-purpose Framework Now, we define our framework, as shown in Figure 1. This framework is based on the design principles discussed above and supports the operations for data selection, temporal aggregation, context filtering, feature vector generation and clustering algorithm selection. Let $\mathcal{D} = \{d_1, \ldots, d_n, d_{n+1}, \ldots, d_{n+m}\}$ be a set of sensor devices available, where:

- $\{d_1, \ldots, d_n\}$ is the set of sensors that is the main subject of our analytics task, or *target sensors*, and
- $\{d_{n+1}, \ldots, d_{n+m}\}$ is the set of additional *context sensors*,

In our case, smart meter is an example of a target sensor, whereas temperature, motion, or sound sensors are examples of context sensors.

For a vector $\mathbf{V}$, we write $\mathbf{V}(i)$ to address the i-th element of $\mathbf{V}$. Let $\gamma^{\star}()$ represents application of a specific design principle $\star$ (from R1 to R5) to the input set.

DEFINITION 3.1. (MEASUREMENTS) *We define a* measurement-tuple *as* $s = (t_s, \mathbf{V}_s)$*, where:*

- t_s *is a* timestamp*,*
- $\mathbf{V}_s \in \mathbb{R}^{n+m}$ *is a vector of* sensor values*, i.e.,* $\mathbf{V}_s(i)$ *is the value of* d_i *at time* t_s*.*

A time series of measurement *is defined as* $S = \{s_1, \ldots, s_{|S|}\}$*, where* $s \in S$ *is a measurement-tuple and whenever* $i < j$*, we have* $t_{s_i} < t_{s_j}, \forall 1 \leq i, j \leq |S|$*.*

3.2.1 Data Selection Let $\mathcal{X}$ be the set of consumers, ts_{start}, ts_{end} be the starting and the ending timestamp, and td_{start}, td_{end} be the starting and ending *time of day* that we are interested in, as the selection criteria of **R1**. In addition, let $timeOfDay(t)$ denote the time of day of timestamp t, i.e., the hour, minute, second, and millisecond of t.

Let S_x denotes the time series of measurements from consumer x's premise. For a set of consumer $\mathcal{X}$, we define

$\mathbf{S}_{\mathcal{X}} = \{S_x \mid x \in \mathcal{X}\}$. Let $\mathcal{X}^+ \supseteq \mathcal{X}$. Then, data selection over $\mathbf{S}_{\mathcal{X}^+}$ is $\gamma^{R1}(\mathbf{S}_{\mathcal{X}^+}, \mathcal{X}, ts_{start}, ts_{end}, td_{start}, td_{end}) = \{S'_x \mid x \in \mathcal{X}\}$, where $S'_x = \{s_i \mid s_i \in S_x, ts_{start} \leq t_i \leq ts_{end}, td_{start} \leq timeOfDay(t_i) \leq td_{end}\}$.

3.2.2 Temporal Aggregation Let $T = [\underline{T}, \overline{T}]$ be a time interval, where $\underline{T}$ and $\overline{T}$ both are timestamps as the lower and upper bounds of the interval, respectively. Let $\mathcal{T} = \{T_1, \ldots, T_{|\mathcal{T}|}\}$ be a set of time intervals denoting the temporal aggregation that we are interested in **R2**. For a time series of measurements S, temporal aggregation by $\mathcal{T}$ over S is defined as $\gamma^{R_2}(S, \mathcal{T}) = \{\hat{s}_1, \ldots, \hat{s}_{|\mathcal{T}|}\}$, where $\hat{s}_i = (T_{\hat{s}_i}, \mathbf{V}_{\hat{s}_i})$ is an *aggregated measurement*, $T_{\hat{s}_i} = T_i$ and

$$\mathbf{V}_{\hat{s}_i} = \sum_{t_s \in T_i} \mathbf{V}_s, \forall 1 \leq i \leq |\mathcal{T}|. \tag{3.1}$$

Note that in the Eq. (3.1) above we aggregate by summing up the sensor values. Depending on the application scenario, other aggregation function such as taking the average, maximum, or minimum values can also be used.

Example. For monthly aggregation over one year data (from January to December), we have $|\mathcal{T}| = 12$, where each $T \in \mathcal{T}$ is a one month time interval. Thus, aggregation by $\mathcal{T}$ over any time series S results in $|\gamma^{R_2}(S, \mathcal{T})| = 12$, where each element, i.e., an aggregated measurement, contains the aggregation of sensor values over one month.

3.2.3 Context Filtering We define two context types that can be defined by the user with respect to **R3**, namely *calendar context*, and *measured context*. Calendar context is defined on timestamps, such as summer, January, weekday, or weekend. Measured context is defined on sensor values, such as temperature between 30 and 35 degree, humidity between 50% and 60%.

DEFINITION 3.2. (CALENDAR CONTEXT) *We define a* calendar context u *as a function* $f_u : t \rightarrow \{0, 1\}$*, where* t *is a timestamp. We have* $f_u(t) = 1$*, if* t *belongs to context* u*, and* 0 *otherwise. Let* U *be a set of calendar contexts, a time interval* $T = [\underline{T}, \overline{T}]$ satisfies U *iff* $f_u(\underline{T}) = 1$ *and* $f_u(\overline{T}) = 1, \forall u \in U$.

Example. Let $U = \{summer, weekend\}$ be the set of calendar contexts that we are interested in. We have:

- $f_{summer} : t \rightarrow \{0, 1\}$, which return 1 if t is in summer, and 0 otherwise,
- $f_{weekend} : t \rightarrow \{0, 1\}$, which return 1 if t is on weekend days, and 0 otherwise.

That is, time intervals which satisfies U are intervals whose lower and upper bounds are both in summer and on weekend days.

DEFINITION 3.3. (MEASURED CONTEXT) *We define a* measured context *as a tuple* $w = (\delta_w, \mathcal{X}_w)$ *where* $\delta_w \in \{n+1, \ldots, n+m\}$ *is a sensor index,* $d_{\delta_w} \in D$ *is the context sensor, and* $\mathcal{X}_w$ *is the accepted interval of the values of context sensor* d_{δ_w}*. A sensor values* $\mathbf{V} \in \mathbb{R}^{n+m}$ satisfies *a set of measured context* W *iff* $\mathbf{V}(\delta_w) \in \mathcal{X}_w, \forall w \in W$,

DEFINITION 3.4. (CONTEXT FILTERING) *Let* $\hat{S}$ *be a time series of aggregated measurements,* U *be a set of calendar contexts, and* W *be a set of measured contexts. Context filtering over* $\hat{S}$ *by* U *and* W *is defined as* $\gamma^{R3}(\hat{S}, U, W) = \{\hat{s} \mid \hat{s} \in \hat{S},\ T_{\hat{s}}$ *satisfies* U*, and* $\mathbf{V}_{\hat{s}}$ *satisfies* $W\}$.

3.2.4 Feature Vector Generation In the context of energy consumption sensing, or environmental sensing in general, measurements from different time of day can be very different and hence, it is considered as an important feature for various data mining task. We take that knowledge into account by allowing the features to be built around different time intervals.

Let Γ be a set of time interval sets, where each $\mathcal{T}_i \in \Gamma$ is a time interval set. Let $\phi : 2^{\mathbb{R}} \rightarrow \mathbb{R}$ be a feature builder function (or feature function) with respect to **R4**, which typically is a statistical function, such as mean, median, standard deviation or inter-quartile range. And, let $\hat{S}$ be time series of aggregated measurements, where $\hat{s} = (T_{\hat{s}}, \mathbf{V}_{\hat{s}}), \forall \hat{s} \in \hat{S}$. A feature vectors computed from $\hat{S}$ using ϕ over Γ is $\gamma^{R4}(\hat{S}, \phi, \Gamma) = \mathbf{F} \in \mathbb{R}^{|\Gamma| \cdot n}$, where:

$$\mathbf{F}(i \cdot n + j) = \phi\big(\{\mathbf{V}_{\hat{s}}(j) \mid T_{\hat{s}} \in \mathcal{T}_i, \hat{s} \in \hat{S}\}\big), \quad i = 0, \ldots, |\Gamma| - 1,\ j = 1, \ldots, n. \tag{3.2}$$

The $(i \cdot n + j)$-th element of feature vector $\mathbf{F}$ is computed using function ϕ over the set of aggregated sensor values $\mathbf{V}_{\hat{s}}$ that belong the same time interval set $\mathcal{T}_i$, and target sensor $d_j \in \{d_1, \ldots d_n\}$.

Example. We give an example of hourly features generation. Let us assume that for the data selection, the user is interested in the data of year 2010. For the temporal aggregation, she is interested in hourly temporal aggregation. And, to simplify our example, let us assume that she is not interested in any context filtering, i.e., $U = \{\}$ and $W = \{\}$. Furthermore, smart meter is our only target sensor, i.e., $n = 1$. Assume that we have a time series of aggregated measurement, $\hat{S}$, for the whole year of 2010, where each $\hat{s} \in \hat{S}$ is an (hourly) aggregated measurement for each hour in 2010. Let $\mathcal{T} = \{T_{\hat{s}} \mid \hat{s} \in \hat{S}\}$ be a set of all time intervals in $\hat{S}$, thus we have $|\mathcal{T}| = 365 \cdot 24 = 8760$. Furthermore, let $\Gamma = \{T_1, \ldots, T_{24}\}$ be a set of time interval set, where each time interval set $T_i \subset \mathcal{T}$ contains the time intervals which accounts only for hour i. Let ϕ be a function that calculates mean. Hence, the result of $\gamma^{R4}(\hat{S}, \phi, \Gamma)$ is $\mathbf{F} \in \mathbb{R}^{24}$, where each $\mathbf{F}(i)$ is the mean of hourly consumption of hour i throughout the year 2010.

Generation from a set of context sets and feature functions. There could be a case where we would like to have a feature vector which is a combined result of applying **R3** using a set of context sets and **R4** using a set of feature functions. For example, instead of using mean as the only feature function, we might want to use both mean and median to have a more robust segmentation. Let $\hat{S}$ be the aggregated

measurement satisfying **R1** and **R2**, Θ be the set of context sets, and Φ be the set of feature functions. In addition, let Γ be the set of time interval sets to build the features. For each context set $(U_i, W_i) \in \Theta$ and feature function $\phi_j \in \Phi$, we compute $\mathbf{F_{ij}} = \gamma^{R4}(\gamma^{R3}(\hat{S}, U_i, W_i), \phi_j, \Gamma)$. Finally, we append $\mathbf{F_{ij}}$, one after the other to form the combined feature vector, where $1 \leq i \leq |\Theta|$ and $1 \leq j \leq \Phi$.

3.2.5 Clustering Algorithm Application Given the expert knowledge from the user to apply a specific algorithm, $A \in \mathcal{A}$, and its parameter setting ψ, then our framework should be able to apply it to the consumers' feature vector (**R5** principle).

Let $\mathcal{X}$ be a set of consumers, and $\mathbf{F}_{\mathcal{X}} = \{\mathbf{F}_x \mid x \in \mathcal{X}\}$ be a set of feature vectors of all consumers in $\mathcal{X}$. Then, the application of clustering algorithm A with parameter setting ψ over $\mathbf{F}_{\mathcal{X}}$ results in a set of clusters (or *cluster configuration*, or *configuration*), i.e., $\gamma^{R5}(A, \psi, \mathbf{F}_{\mathcal{X}}) = \{c_1, \ldots, c_k\}$, where each cluster $c_j \subseteq \mathcal{X}$, $\forall 1 \leq j \leq k$, is a set of consumers.

Automatic Cluster Configuration Selection. Given different parameter settings, we are often uncertain which parameter setting delivers us the best cluster configuration (according to some cluster evaluation metrics). This holds even if the setting is simple and easy to understand, such as the number of clusters to be created (in case of using k-means algorithm). For instance, we are often uncertain in choosing the value of k, the number of clusters. This motivates us to include an automatic selection of cluster configuration in our framework. Our selection mechanism is similar to the mechanism in [14, 24], i.e., we attempt to select compact and well separated clusters.

Given a clustering algorithm A, a set of parameter settings $\Psi = \{\psi_1, \ldots, \psi_{|\Psi|}\}$, and consumers' feature vector $\mathbf{F}_{\mathcal{X}}$, we can have a set of cluster configuration $\mathcal{C} = \{C_i \mid \gamma^{R5}(A, \psi_i, \mathbf{F}_{\mathcal{X}}) = C_i\}$. Thus, our task is to select the best cluster configuration $C^* \in \mathcal{C}$. In order to determine C^*, we use three cluster evaluation metrics: Silhouette index [22], Dunn index [9], and Davies-Bouldin index [8].[1] We perform a majority voting to the best configuration identified by each metrics. Algorithm 3.1 describes the selection mechanism in more details.

Functions $sortSilhouette(\mathcal{C})$, $sortDunn(\mathcal{C})$, and $sortDaviesBouldin(\mathcal{C})$ compute Silhouette, Dunn, and Davies-Bouldin index respectively for each configuration in $\mathcal{C}$, and return an ordered list of the cluster configurations, sorted by the configuration quality in descending order (that is, sorted in decreasing order for Silhouette and Dunn indices, and in increasing order for Davies-Bouldin index). Let $count(L, e)$ be the count of element e in the list L. Function $mostFrequent(L)$ returns an element e^* in L, where $count(L, e^*) > count(L, e)$ for all $e \neq e^*$ in L. In other words, $mostFrequent(L)$ returns the element with the largest count, e^*, and there is no other element in L which has the same count as e^*. If there is no such element, this function returns **null**.

[1] See more detailed description in [1].

Algorithm 3.1: Automatic Cluster Configuration Selection

Input: a set of cluster configuration $\mathcal{C} = \{C_1, \ldots, C_{|\mathcal{C}|}\}$
Output: the best configuration $C^* \in \mathcal{C}$

```
silhList ← sortSilhouette(C)
dunnList ← sortDunn(C)
davbList ← sortDaviesBouldin(C)
countList ← [ ]
C* ← null
i ← 1
repeat
    countList.add(silhList[i])
    countList.add(dunnList[i])
    countList.add(davbList[i])
    C* ← mostFrequent(countList)
    i ← i + 1
until (i > |C|) ∨ (C* = null)
return C*
```

4 Clustering Consistency

This section answers two important questions. First, *is there any consumer who changes their behavior over time?* For example, we would like to know whether there is a consumer who is in the low consumption cluster in January, but she changes to the medium/high consumption cluster in February. This insight is important for devising personalized feedback to the consumer. Second, *how different is one cluster configuration to another?* For example, how different is the cluster configuration using January data compared to using February data, or March, April, etc. Answering this question gives insight to the utility company on the key contexts to consider when developing policies, such as differential pricing or demand response signal. In the sequel, we use the term *individual* and *consumer* interchangeably.

4.1 Individual to Cluster Consistency Given two cluster configurations, we develop a measure to indicate whether an individual has the same cluster membership on both of them. Because cluster configuration is invariant to the cluster labels, we require the measure to also be invariant to the label sets. Thus, the idea is to define an *individual to cluster consistency* index ($i2c$) which computes how consistent are the fellow cluster members of an individual on the two configurations.

Let C be a cluster configuration. We write $C(x)$ to denote the cluster of x, i.e., the cluster $c \in C$ where $x \in c$. We define an individual to cluster consistency of consumer $x \in \mathcal{X}$ on two cluster configurations C_1 and C_2 as:

$$i2c(x, C_1, C_2) = \frac{|C_1(x) \cap C_2(x)| + |(\mathcal{X} \setminus C_1(x)) \cap (\mathcal{X} \setminus C_2(x))| - 1}{|\mathcal{X}| - 1}. \tag{4.3}$$

Intuitively, if we denote the set of consumers in $C(x)$ as the *friends* of x, and the others as *non-friends*, then $i2c$ measure

the number of friends in C_1 who also friends of x in C_2, and the number of non-friends in C_1 who also non-friends in C_2 normalized by the number of all consumers excluding x. The value of $i2c$ ranges between 0 and 1. The closer it is to 1, the more consistent is x's cluster membership in C_1 and C_2.

4.2 Distance Rank Given individuals who change their clusters, one might interested more to the ones located closer the centroid of their new clusters. Thus, we define an additional measure, *distance rank*, to denote the ranking of individual's distance to its cluster representative (such as centroid) compared to the other cluster members. We use distance rank instead of actual distance measure because it is invariant to cluster size. Thus, it can be used for comparison across different clusters.

Let $C(x)$ be the cluster of x in configuration C, and $\zeta^{C(x)}$ be the cluster representative of $C(x)$. In addition, let $dist(x, \zeta^{C(x)})$ be the distance of x to its cluster representative. For an individual x, we define its distance rank as:

$$dr(x, C(x)) = \frac{|\{x' \mid dist(x, \zeta^{C(x)}) < dist(x', \zeta^{C(x)}), x' \in C(x)\}|}{|C(x)|}. \tag{4.4}$$

Distance rank ranges between 0 and 1. The higher the value, the closer the individual to its cluster representative.

4.3 Cluster Configuration Consistency In order to investigate community behavioral changes over different contexts, we can measure the difference of the resulting cluster configuration over those contexts. Let $\mathcal{X}$ be a set of consumers, and C_1 and C_2 be two cluster configurations over $\mathcal{X}$. We compute the difference between C_1 and C_2, i.e., *cluster configuration consistency* index, as the average of $i2c$ of their individuals:

$$ccc(C_1, C_2, \mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} i2c(x, C_1, C_2). \tag{4.5}$$

The ccc index ranges between 0 and 1. The higher the ccc between C_1 and C_2, the more similar they are.

5 Knowledge Extraction from Survey Data

Consumer segments can be useful for implementing different policies, such as: targeted demand responses, more personalized energy feedback, or differential pricing. However, having consumer segmentation alone is not enough. We also need to understand *what are the characteristics that constitute a consumer cluster*? Only by developing this understanding, we can develop an effective and efficient policies which tailored better for our consumers.

Consumer characteristics can be of form demographic profiles, house types, appliance usages, or living styles. We can obtain these through survey/questionnaire, using questions, such as: What best describes the people you live with (single/adults/adults with children)? Do you have a dishwasher (yes/no)? What is the approximate floor area of your home?[2] In this section, we focus on mining consumers characteristics, which are the discriminative, i.e., characteristics which make a cluster different from the others. We model consumer characteristic as a pair of question and answer. Next, we describe how to compute *discriminative index*.

5.1 Discriminative Index We define a measure to express how discriminative a question-answer pair is in distinguishing a cluster from the others. Let $\mathcal{X}$ be a set of consumers, and C be a cluster configuration over $\mathcal{X}$. For a cluster $c \in C$, we denote $\neg c$ as all individuals who are not in c, i.e., $\neg c = \{x' \mid x' \in \mathcal{X} \setminus c\}$. In addition, let q be a question, $ans(q)$ be the set of possible answers to q, $ans_x(q) \in ans(q)$ be the answer of consumer x to question q, and $N_{c,q}$ be the set of consumers in cluster c who respond to question q.

We define $Z_c(q, a)$ as the fraction of individuals in cluster c who answer a to question q, i.e.,

$$Z_c(q, a) = \frac{|\{x \mid ans_x(q) = a,\ x \in c\}|}{|N_{c,q}|}, \tag{5.6}$$

where $a \in ans(q)$. Then, *discriminative index* of question q and answer a to cluster c is defined as:

$$DI_c(q, a) = \frac{Z_c(q, a) - Z_{\neg c}(q, a)}{\max\left(Z_c(q, a), Z_{\neg c}(q, a)\right)}. \tag{5.7}$$

Discriminative index ranges between -1 and $+1$. It is discriminative positive (or negative) if it is positive (or negative). Both discriminative positive and negative explain how a cluster differs from the others. Discriminative index close to $+1$ means that most of the individuals in cluster c answer a to question q, whereas individuals in other clusters do not. In contrast, discriminative index close to -1 means that most of individuals in other clusters answer a to question q, whereas individuals in cluster c do not. Discriminative index close to 0, means that answer a to question q does not differentiate cluster c from the others, i.e., it has little or no discriminative power for cluster c.

5.2 Dealing with Ordinal and Quantitative Data In a survey, there are some answers which are ordinals or quantitatives. For example, in our survey data, answers to the question whether a consumer would like to do more to reduce electricity usage, are ordinals, i.e., five criteria from strongly agree to strongly disagree. Another example: answers to a question of the approximate floor area of the house, are quantitative, i.e, the number which represent the floor area.

In the previous section, we determine whether a specific pair of question and answer is a key characteristic of a cluster. However in ordinal and quantitative answers, we are interested on insights more than that. For example, instead of "most of the consumers in cluster c have *X* sq ft. floor area ", we are more interested on more general insight, if any, such as: "most of the consumers in cluster c have *less than X* sq ft.

[2] These example questions are taken from our dataset (explained in Section 6).

floor area", or "*between X and Y*", or "*greater than X*". One way to do this is by introducing some splitting points which divide the answers into groups, or ranges. But then, it leads us into a combinatorial problem, such as: how many points do we need for the best splitting, and where should we put the splitting points.

Let q be a question, and $ans(q)$ be a set of consumers' answers to q. To solve this problem, we sort the answers in ascending order and put them into a list. We add $-\infty$ and $+\infty$ to the beginning and the end of the list. Let $l = |ans(q)| + 2$ be the length of the list. We take all possible n-grams from the list, where n varies from $l-1$ to 1. Next, we create a set of ranges $R_{ans(q)}$ by taking the first and the last element of each n-grams as the lower and upper bounds (inclusive). Finally, we remove ranges $[-\infty, -\infty]$ and $[\infty, \infty]$ from $R_{ans}(q)$. This takes polynomial time on the size of $ans(q)$.

Example. Let $ans(q) = \{1, 2, 5\}$. The set of ranges created from all possible n-grams from length 4 to 1, without $[-\infty, -\infty]$ and $[\infty, \infty]$, is $R_{ans(q)} = \{\ [-\infty, 5]$, $[1, \infty]$, $[-\infty, 2]$, $[1, 5]$, $[2, \infty]$, $[-\infty, 1]$, $[1, 2]$, $[2, 5]$, $[5, \infty]$, $[1, 1]$, $[2, 2]$, $[5, 5]\ \}$.

Then, for ordinal and quantitative questions/answers, instead of computing discriminative indices based on $a \in ans(q)$, we now compute them based on $a^r \in R_{ans(q)}$:

$$Z_c(q, a^r) = \frac{|\{x \mid ans_x(q) \in a^r,\ x \in c\}|}{|N_{c,q}|}, \tag{5.8}$$

$$DI_c(q, a^r) = \frac{Z_c(q, a^r) - Z_{\neg c}(q, a^r)}{\max\left(Z_c(q, a^r), Z_{\neg c}(q, a^r)\right)}. \tag{5.9}$$

6 Experimental Evaluations

In this section, we describe our experiment details. We use Irish CER dataset, which contains energy consumption measurements of around 5,000 consumers over 1.5 years [2]. The measurements started in July 2009 and ended in December 2010, and recorded energy consumption in kWh every 30 minutes. We choose residential consumers that belong to the control group and have no missing values. This results in the selection of 782 consumers. Smart meter is the only target sensor, $n = 1$, and there is no context sensor, $m = 0$. In addition, the dataset also contains survey results, which includes consumers' demographics (occupation, family type, etc.), house information (ownership, age, floor area, etc.), and appliance usages (dishwasher, TV, water pump etc.).[3]

6.1 Consumer Segmentation First, we demonstrate the result of our selection mechanism (described in Section 3.2.5). Second, we show the result of our framework to accomplish different consumer segmentation tasks, i.e., consumer segmentation by consumption trends, absolute consumption, and consumption variability. We also show that applying the same task on different contexts yield different results. We visualize each cluster by its centroids.

[3] http://www.ucd.ie/issda/static/documentation/cer/smartmeter/cer-residential-pre-trial-survey.pdf

6.1.1 Automatic Cluster Configuration Selection Our automatic selection mechanism aims to select compact clusters which far from each other (well separated). We perform a consumer segmentation task based on their consumption trends. We select all data (R1). We use hourly temporal aggregation (R2). Our features is a combined feature vector, from a set of context sets (R3) and feature functions (R4). We use calendar context only (without measured context), i.e., the set of context sets $\{(U, W)\}$ is $\big\{\big(\{$January, weekday$\}, \{\}\big), \big(\{$January, weekend$\}, \{\}\big)\big\}$. To obtain the consumption trends, we use {normalized mean, normalized median} as the set of feature functions. *Normalized* here means that we apply standard normalization on the measurement tuple, such that its mean is 0 and its standard deviation is 1. We use k-means algorithm, for $k = 2, \ldots, 10$ (R5). As a consequence, we obtained 9 different cluster configurations.

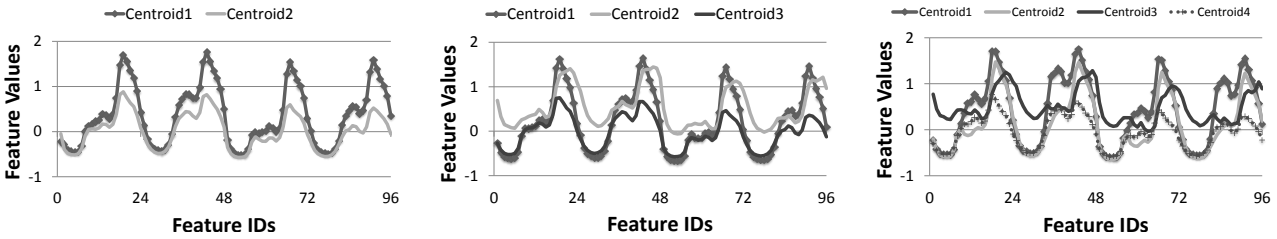
Cluster configuration resulting from $k = 2$ is determined as the best configuration by our automatic selection mechanism. Figure 2 illustrates the result using $k = 2$, $k = 3$, and $k = 4$. We can see that centroids of clusters using $k = 2$ are better separated than the others. In addition, the configuration separates the consumers who have high and low peak consumption in the evening.

6.1.2 Various Segmentation Tasks In this section, we show the generality of our framework to accomplish different consumer segmentation tasks. Furthermore, we show that applying consumer segmentation in different contexts produce different results.

We perform consumer segmentation by consumption trends, absolute consumption, and consumption variability. The setting are similar as in Section 6.1.1, except for the set of context sets and feature functions. We use the set of feature functions {normalized mean, normalized median} for consumption trends, {mean, median} for absolute consumption, and {standard deviation, IQR} for consumption variability. We perform the task in three different contexts: January, July, and all months, and separate weekend consumption from weekdays. Thus, we use the set of context sets: $\big\{\big(\{$January, weekday$\}, \{\}\big), \big(\{$January, weekend$\}, \{\}\big)\big\}$ for January, $\big\{\big(\{$July, weekday$\}, \{\}\big), \big(\{$July, weekend$\}, \{\}\big)\big\}$ for July, and $\big\{\big(\{$weekday$\},\{\}\big), \big(\{$weekend$\}, \{\}\big)\big\}$ for all months.

Figure 3 shows that for segmentation by consumption trends, we successfully divide the consumers into high peak and low peak consumers. For the next tasks, segmentation by absolute consumption and consumption variability, we are also able to produce clusters with high and low absolute consumption and consumption variability, respectively. Furthermore, comparing the results on January, July, and all months, shows that applying the segmentation tasks in different contexts yields different results. This validates our approach to perform context-based consumer segmentation.

6.2 Clustering Consistency Using our clustering consistency index (described in Section 4), we can quantify the difference between cluster configurations. Figure 4a shows



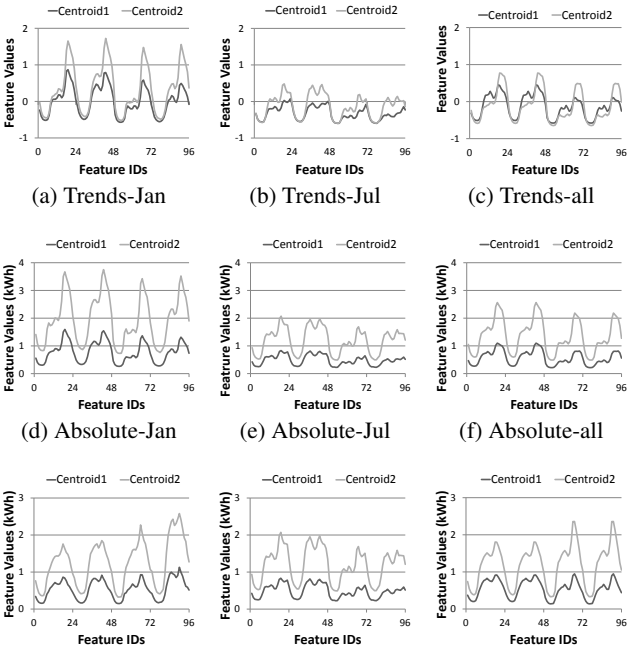
(a) $k = 2$ (b) $k = 3$ (c) $k = 4$

Figure 2: Centroids of clusters using January data, and hourly temporal aggregation. For the features, we use normalized mean of weekday (ID 1-24), normalized mean of weekend (ID 25-48), normalized median of weekday (ID 49-72), and normalized median of weekend (ID 73-96) consumption. We use k-means algorithm with $k = 2$, $k = 3$, and $k = 4$.

the difference between cluster configurations, resulting from consumer segmentation by absolute consumption for a specific month compared to 1, 3, and 6 months previously. We use k-means algorithm, with k=2. This results in two consumer segments: high and low consumption clusters.

The result shows that the consistency between clusters resulting from the segmentation of the current month and 1 month ago is higher than the consistency between clusters from the current month and 3 months or 6 months ago. Especially, the lowest consistency is between July 2010 and 6 months previously, January 2010 (cluster configuration consistency = 0.67). One of the reason is seasonal changes in the energy consumption behavior between summer and winter, i.e., July and January is in the middle of summer and and winter in Ireland, respectively. However the implication of our result could be bigger than that. It indicates that, there are a number of consumers who behave differently (compared to their fellow cluster members), which in turn change their cluster memberships.

To elaborate this, in Figure 4b and 4c, we show the consumption profile (mean and median of weekday and weekend consumption) of the centroid of the low consumption cluster, and two individuals: ID 1301 and ID 7381, in January and July 2010. These two consumers are in the low consumption clusters in January 2010. Both of them have the highest distance rank among individuals with low consistency index (ID 1301) and high consistency index (ID 7381) between January and July 2010. Consumer ID 1301 changes her cluster membership from low consumption cluster in January 2010 to the high consumption cluster in July 2010 (we use centroid as the cluster representative). The typical consumption of the low consumption cluster in July 2010 is lower than January 2010, which shows the seasonal changes in the electricity consumption between winter and summer. Consumer ID 7381, who stays in the low consumption cluster, lower her consumption level, in line with the behavior of her cluster. However, the consumption of consumer ID 1301 in July 2010 is approximately the same as her consumption in January 2010. This causes her to change her cluster membership to the high consumption cluster in July 2010. Then, given this results, we could devise a personalized energy feedback for consumer ID 1301 to lower her energy consumption (for example, using self-comparison).



(a) Trends-Jan (b) Trends-Jul (c) Trends-all

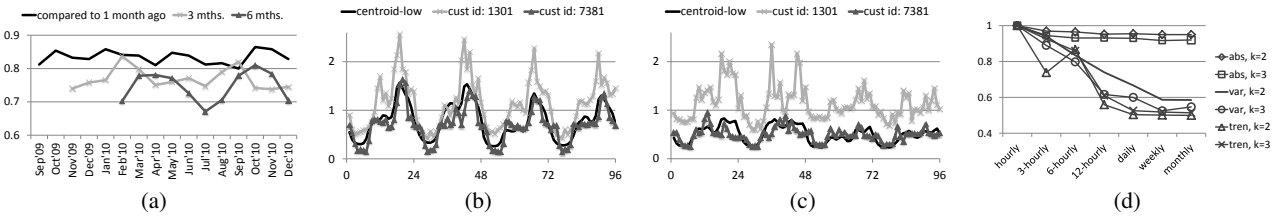
(d) Absolute-Jan (e) Absolute-Jul (f) Absolute-all

(g) Variability-Jan (h) Variability-Jul (i) Variability-all

Figure 3: Consumer segmentation on consumption trends (a)-(c), absolute consumption (d)-(f), and consumption variability (g)-(i), in different contexts: January, July, and all months. For trends, we used the same features as in Figure 2. Feature ID 1-24 and 49-72 are weekdays consumption, whereas 25-48 and 73-96 are weekend consumption. We use mean (ID 1-48) and median (ID 49-96) for absolute, and standard deviation and IQR for variability.

In addition, our cluster configuration consistency index is useful to quantify the difference between results obtained by various temporal aggregations. Since smart meter data has high velocity and high volume, determining the right aggregation is imperative. We compare the segmentation results by absolute consumption, consumption variability, and trends, performed using different temporal aggregations, against hourly temporal aggregation.

Figure 4d shows that, for consumer segmentation by absolute consumption, we have only a little difference between cluster configurations resulting from various temporal aggre-



(a) (b) (c) (d)

Figure 4: (a) cluster configuration consistency over time (monthly), consumption profile using (b) January 2010 data, (c) July 2010 data, and (d) cluster configuration consistency over different temporal aggregations.

Table 1: Consumer characteristics based on their absolute consumption. A minus (-) sign denotes discriminative negative.

Cluster	Question	Answer	DI
low	family type	single	0.86
	floor area (sq ft.)	805-1073	0.86
	#bedrooms	≤ 2	0.85
medium	electric shower	(-) $\geq$ 20 mins	-0.76
	family type	(-) single	-0.61
	floor area (sq ft.)	2300-2750	0.56
high	#children	≥ 4	0.93
	family type	(-) single	-0.90
	floor area (sq ft.)	(-) $\leq$ 1200	-0.87

Table 2: Consumer characteristics based on their consumption variability. A minus (-) sign denotes discriminative negative.

Cluster	Question	Answer	DI
low	water pump	(-) 1-2hrs	-0.88
	family type	single	0.80
	washing machine	(-) $\geq$ 2-3 loads	-0.76
medium	electric shower	10-20 mins	0.59
	family type	(-) single	-0.55
	#children	(-) ≥ 3	-0.54
high	tumble dryer	$\geq$ 2 to 3 loads	0.90
	#children	≥ 4	0.88
	floor area (sq ft.)	$\geq$ 2800	0.79

gations and hourly aggregation. This is a good news because storing and processing monthly aggregated data, for example, is far more desirable than hourly data. Unfortunately, for consumer segmentation by consumption variability and trends, this is not the case. The consistency index decreases as the temporal aggregation become coarser, i.e., the coarser the temporal aggregation, the higher the difference with the hourly aggregation. This can be understood since as we move to coarser temporal aggregation we lose the variation in the consumption profiles, which is needed to distinguish different consumption variability or trends.

6.3 Knowledge Extraction from Survey Data Our dataset contains not only smart meter measurement, but also consumer survey data (questions/answers). From this survey data, using our discriminative index (described in Section 5) we are able to extract knowledge about main characteristics of a cluster. This step is imperative for applying the right business decision or policy to a specific consumer segment.

In Table 1 and 2, we show the top 3 characteristics of consumer segments, formed by absolute consumption and consumption variability, respectively. We use the same features and setting as in Figure 3f and 3i, with k=3 (number of clusters). The characteristics (expressed by questions and answers) are ordered by the absolute value of their discriminative index (DI). Recall that DI < 0 denotes discriminative negative, i.e., the answer is associated more likely to other clusters.

Consumer segments by absolute consumption is determined more by consumer's demographics (such as family type, the number of children) and housing condition (floor area, the number of bedrooms). Low consumption cluster is dominated by single, whereas medium/high consumption clusters are dominated by non-single family, either adults only or adults with children. Floor area is also relevant, with low consumption clusters having the smaller area.[4]

There are more insights which are based on appliance usages in Table 2. It can be understood since consumption variability comes from intermittent appliance usages. Consumers with low consumption variability use big appliances, such as water pump and washing machine, in a shorter duration.[5] Furthermore, consumers with high consumption variability use tumble dryers for the longest duration.

Often opinion about house's energy consumption is built around its floor area or the year it was built.[6] While Table 1 and 2 shows insights about the floor area, there is none about the year it was built. To investigate this further, we plot the cumulative distribution of consumers' floor area (Figure 5a) and the year their houses was built (Figure 5b). Figure 5a shows that, indeed floor area is a relevant characteristics, i.e., we can distinguish clearly the cumulative distribution of the floor area among different clusters, where consumers in the lower consumption cluster typically associated with smaller floor area. Figure 5b, however, shows that this is not the case with the year the houses was built, where the cumulative distribution for the three clusters are similar (coincide) to each other.

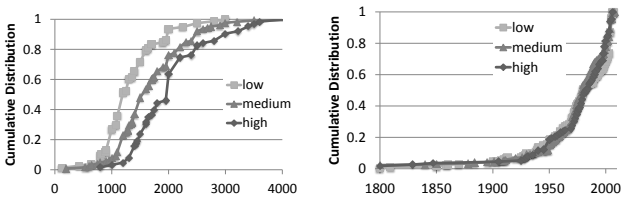
7 Conclusion

In this paper, we presented a generic consumer segmentation framework that can be used to classify smart meter data into clusters using multiple distinguishing characteristics such

[4] Maximum consumers' floor area is around 5000 sq ft.

[5] Water pump usages are ranges from $<$30 minutes to $>$2 hours. Washing machine usages are ranges from $<$1 load to $>$3 loads a day.

[6] In [1], we also investigate whether ownership of a certain appliance is discriminative to the consumer clusters.



(a) Floor area (sq ft.) (b) Year built

Figure 5: Cumulative distribution of (a) floor area, and (b) the year consumers' houses was built. Consumers are clustered by their absolute consumption: low, medium, and high.

as time of consumption, levels of consumption, associated contexts, etc. We also presented a clustering consistency index, which can be used to track evolving consumption behaviors and to compare consumer segments resulting from different temporal aggregations.

We evaluated the framework and index using real world smart meter data and survey results. Our experiments showed that consumer segmentation results are different from one context to another. Moreover, different temporal aggregations have only a little effect on segmentation by absolute consumption. But, this does not hold for segmentation by consumption variability or trends. Furthermore, consumer's floor area is relevant to her consumption. In addition, big appliances' usage patterns also play a role in consumer's consumption variability.

In the future, we plan to define consumer segments based on their responses to DR signals. We aim to quantify and understand consumer characteristics which are relevant to the responses. This enables us to develop DR design recommendation framework, which allows us to draw relationships between consumers' consumption profile, demographics, and appliance usage patterns to estimate the impact of future (planned) DR signals. The existence of this framework is crucial for delivering effective DR signals. In addition, we also plan to explore applications of our framework in anomaly, theft, and fraud detection.

References

[1] Supplementary material. `https://github.com/tritritri/consumer-segmentation`.

[2] Electricity customer behaviour trial. The Commission for Energy Regulation (CER), 2012.

[3] A. Albert and R. Rajagopal. Smart meter driven segmentation: What your consumption says about you. *IEEE Transactions on Power Systems*, PP(99):1–12, 2013.

[4] Beyond Zero Emissions. Smart Meters are a Smart Choice. `http://engage.haveyoursay.nsw.gov.au/document/show/748`, 2011.

[5] C.-S. Chen, J. C. Hwang, and C. Huang. Application of load survey systems to proper tariff design. *IEEE Transactions on Power Systems*, 12(4):1746–1751, 1997.

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

[7] G. Chicco, R. Napoli, P. Postolache, M. Scutariu, and C. Toader. Customer characterization options for improving the tariff offer. *IEEE Transactions on Power Systems*, 18(1):381–387, 2003.

[8] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[9] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[10] V. Figueiredo, F. Rodrigues, Z. Vale, and J. Gouveia. An electric energy consumer characterization framework based on data mining techniques. *IEEE Transactions on Power Systems*, 20(2):596–602, 2005.

[11] C. Flath, D. Nicolay, T. Conte, C. Dinther, and L. Filipova-Neumann. Cluster analysis of smart metering data. *Business & Information Systems Engineering*, 4(1):31–39, 2012.

[12] F. Fusco, M. Wurst, and J. W. Yoon. Mining residential household information from low-resolution smart meter data. In *21st International Conference on Pattern Recognition (ICPR)*, pages 3545–3548, 2012.

[13] M. Kitayama, R. Matsubara, and Y. Izui. Application of data mining to customer profile analysis in the power electric industry. In *IEEE Power Engineering Society Winter Meeting, 2002*, volume 1, pages 632–634 vol.1, 2002.

[14] F. Martinez Alvarez, A. Troncoso, J. Riquelme, and J. Aguilar Ruiz. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1230–1243, 2011.

[15] S. J. Moss, M. Cubed, and K. Fleisher. Market segmentation and energy efficiency program design. *California Institute for Energy and Environment Berkeley*, 2008.

[16] ORACLE. Smart Metering for Electric and Gas Utilities. `http://www.oracle.com/us/industries/utilities/046593.pdf`, 2011.

[17] P. Durand. Market Segmentation: Defining A New Attitude For Utilities. `http://bit.ly/vwbEiF`, 2011.

[18] M. Pedersen. Segmenting residential customers: Energy and conservation behaviors. *ACEEE Summer Study on Energy Efficiency in Buildings*, 2008.

[19] S. Ramos, Z. Vale, J. Santana, and J. Duarte. Data mining contributions to characterize MV consumers and to improve the suppliers-consumers settlements. In *IEEE Power Engineering Society General Meeting*, pages 1–8, 2007.

[20] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):pp. 846–850, 1971.

[21] T. Räsänen and M. Kolehmainen. Feature-based clustering for electricity use time series data. In *Proceedings of the 9th International Conference on Adaptive and Natural Computing Algorithms*, ICANNGA'09, pages 401–412, Berlin, Heidelberg, 2009. Springer-Verlag.

[22] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.

[23] S. Harrison. Smart Metering Projects Map. `http://bit.ly/GSs1o5`, 2013.

[24] W. Shen, V. Babushkin, Z. Aung, and W. L. Woon. An ensemble model for day-ahead electricity demand time series forecasting. In *Proceedings of the Fourth International Conference on Future Energy Systems*, e-Energy '13, pages 51–62, New York, NY, USA, 2013. ACM.

(Supplementary Material) Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data *

Tri Kurniawan Wijaya † Tanuja Ganu ‡ Dipanjan Chakraborty‡ Karl Aberer†
Deva P. Seetharam‡

1 Supplement for Automatic Cluster Configuration Selection in Section 3.2.5

We give a brief description about the Silhouette [3], Dunn [2], and Davies-Bouldin [1] indices. They provide us a way to compare a cluster configuration from one to another. However, there are some differences.

Let x be a consumer, C be a cluster configuration (set of clusters), and $C(x) \in C$ be the cluster of x. In addition, let $dist(x, x')$ be the distance between two consumers x and x'.

The Silhouette index This index determines how well an object is clustered, based on the difference in the dissimilarity of the object to its cluster and to the other clusters.

Let $dist(x, c)$ be the average distance between x and all consumers in c, i.e.,

$$dist(x, c) = \frac{1}{|c|} \sum_{x' \in c} dist(x, x').$$

Let $a(x)$ be the average dissimilarity of consumer x to all other fellow cluster members in $C(x)$, i.e,

$$a(x) = \frac{1}{|C(x)| - 1} \sum_{\substack{x' \in C(x) \\ x' \neq x}} dist(x, x').$$

Assuming that $dist(x, x) = 0$, then we can also rewrite the equation above into:

$$a(x) = \frac{dist(x, C(x))}{|C(x)| - 1}.$$

Let $b(x)$ be the minimum average dissimilarity between x and other clusters, i.e.,

$$b(x) = \min_{c \neq C(x)} \frac{dist(x, c)}{|c|}.$$

Then, we define the Silhouette value of x as:

$$silh(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

The Silhouette index of a cluster configuration is the average of the Silhouette index of all consumers (in the configuration):

$$silh(C) = \frac{1}{|C|} \sum_{c \in C} \Big(\frac{1}{|c|} \sum_{x \in c} silh(x) \Big)$$

Silhouette index range from -1 to +1. The closer it is to 1, the better.

The Dunn index This index seeks the largest inter-cluster distance and the lowest intra-cluster distance. The Dunn index is computed based on the ratio between the minimum inter-cluster distance and the maximum intra-cluster distance.

Let us define the inter-cluster distance between two clusters, c_1 and c_2, as the minimum distance between any two points in c_1 and c_2, i.e.,

$$interdst(c_1, c_2) = \min_{\substack{x_1 \in c_1 \\ x_2 \in c_2}} dist(x_1, x_2),$$

In addition, we define the intra-cluster distance (or *diameter*) of a cluster c, as the maximum distance between any two points in c, i.e.,

$$dia(c) = \max_{x_1, x_2 \in c} dist(x_1, x_2).$$

Then, we define the Dunn index of a configuration C as:

$$dunn(C) = \frac{\min\limits_{\substack{c_1, c_2 \in C \\ c_1 \neq c_2}} interdst(c_1, c_2)}{\max\limits_{c \in C} dia(c)}.$$

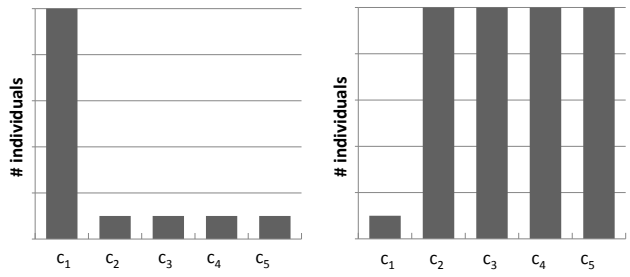
The larger the Dunn index, the better.

The Davies-Bouldin index This index is similar to the Dunn index, i.e., it aims to indentify a cluster configuration which has the largest inter-cluster distance and the lowest intra-cluster distance. The Davies-Bouldin index is computed

*Supported by European Union's Seventh Framework Programme (FP7/2007-2013) 288322, Wattalyst.

†School of Computer and Communication Sciences, EPFL, Switzerland. {tri-kurniawan.wijaya,karl.aberer}@epfl.ch

‡IBM Research India. {tanuja.ganu,cdipanjan,dseetharam}@in.ibm.com



(a) Feature f_1 (b) Feature f_2

Figure 1: Feature f_1 is discriminative positive for cluster c_1, whereas f_2 is discriminative negative for c_1. While entropy measure is able to recognize only discriminative positive features, our *discriminative index* is able to recognize both, discriminative positive and negative features.

based on the sum of diameter between two clusters divided by their inter-cluster distance:

$$daviesBouldin(C) = \frac{1}{|C|} \sum_{c_1 \in C} \max_{\substack{c_2 \in C \\ c_2 \neq c_1}} \frac{dia(c_1) + dia(c_2)}{interdst(c_1, c_2)}$$

In this case, we define the intra-cluster distance of a cluster c as the average distance of the cluster members to its centroid, i.e.,

$$dia(c) = \frac{1}{|c|} \sum_{x \in c} dist(x, \zeta^c),$$

where ζ^c is the centroid of cluster c. We define the inter-cluster distance to be similar with the one used for computing the Dunn index. Note that we define the Davies-Bouldin index here a little bit different compared to its original version [1]. However, as long as $dist$ is a proper distance metric, our definition satisfies Definition 1 to 5 in [1]. The lower the Davies-Bouldin index, the better.

2 Supplement for Section 5

An alternative to discriminative index Entropy can be used as an alternative to our *discriminative index* for determining whether a certain consumer characteristics is discriminative or not, using the same idea as in the decision tree learning. However, there is a subtle difference.

Using entropy, a feature is said to be discriminative for a particular class (or cluster, in our case) when it has low entropy. In Figure 1, f_1 has low entropy, and hence it is discriminative. That is, f_1 is an appropriate feature to distinguish cluster c_1 from others. Moreover, f_1 as an example of what we called as a *discriminative positive* feature. Feature f_2 in Figure 1, has high entropy. Thus, according to the entropy measure, f_2 is not discriminative. However, we can see that f_2 is actually also a discriminative feature, i.e., it characterizes an individual which does not belong to c_1 (might belong to any other clusters). Feature f_2 is an example of what we called as a *discriminative negative* feature.

While entropy is useful measure to recognize discriminative positive feature, it does not recognize discriminative negative feature. Our *discriminative index*, on the other hand, is able to distinguish both, discriminative positive and negative features.

3 Supplement for Section 6.3

Compared to appliance usage, information about appliance ownership is simpler and cheaper to obtain. Using questionnaire is enough to obtain the information whether a consumer own a certain appliance. Detailed appliance usage information, however, is more expensive to obtain because it involves sensor measurement.[1] Thus, knowing whether ownership of a particular appliance determines consumer's energy consumption is a valuable insight.

In our dataset, we have a set of question/answer whether a consumer own these appliances:

- washing machine,
- tumble dryer,
- dishwasher,
- electric shower,
- electric cooker,
- stand alone freezer,
- water pump,
- immersion,
- TV less than 21 inch,
- TV greater than 21 inch,
- desktop computer,
- laptop computer, and
- games consoles.

In Table 1 and 2, we show customer characteristics which related to appliance ownership only. Both shows how discriminative is an ownership of a particular appliance for different clusters, based on absolute consumption and consumption variability. Let $support$ be Z_c in case of discriminative positive and $Z_{\neg c}$ in case of discriminative negative. We show only characteristics with DI ≥ 0.6 (highly discriminative) and $support \geq 0.4$ (highly evident).

[1] Typical appliance usage, however, as in our dataset, can be obtained through questionnaire.

Table 1: Discriminative appliances' ownership for different clusters based on their absolute consumption. We show only for DI $\geq$ 0.60 and support $\geq$ 0.40. A minus (-) sign denotes discriminative negative.

#	Appliance	Cluster	Ownership	DI
1	dishwasher	high	(-) no	-0.76
2	games consoles	low	(-) yes	-0.70
3	tumble dryer	low	no	0.68
4	dishwasher	low	no	0.67
5	games consoles	high	yes	0.61

Table 2: Discriminative appliances' ownership for different clusters based on their consumption variability. We show only for DI $\geq$ 0.60 and support $\geq$ 0.40. A minus (-) sign denotes discriminative negative.

#	Appliance	Cluster	Ownership	DI
1	dishwasher	high	(-) no	-0.72
2	tumble dryer	high	(-) no	-0.72
3	tumble dryer	low	no	0.71
4	games consoles	low	(-) yes	-0.69
5	dishwasher	low	no	0.67
6	games consoles	high	yes	0.60

Over all appliances, we found that, only the ownership of big (power consuming) appliances (dishwasher and tumble dryer), which are highly discriminative. That is, the owner of these appliances are more likely to consume more energy and have higher consumption variability. The ownership of other appliances, which are not shown in Table 1 and 2, are less discriminative.[2]

The consistent presence of games consoles in both tables, however, is rather interesting since they are not big appliances (their power consumption is comparable to other electronic devices such as TV or desktop computer). We conjecture that the ownership of games consoles is highly correlated with family type, e.g., families with children are more likely to have games consoles at home compared to singles. Because family type is a highly discriminative characteristics for households' energy consumption behavior (see Table 1 and 2 in the main paper), then its correlation with games consoles ownership explains why games consoles ownership is also discriminative. Our conjecture is then confirmed in Figure 2, where it shows that, indeed, families with children are the most likely to own games consoles, followed by adults only families, and then by singles, who are the least likely to own games consoles.

[2]However, their usage pattern might be highly discriminative (such as washing machine, electric shower, water pump – see Table 1 and 2 in the main paper).

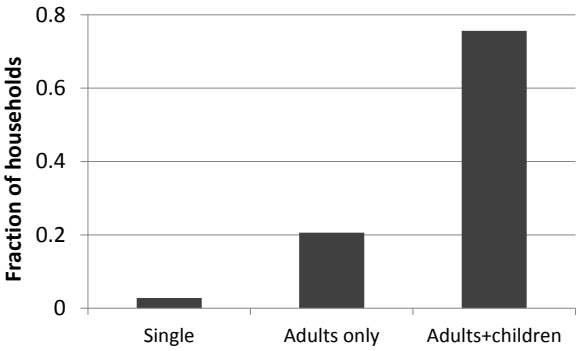


Figure 2: Fraction of households which own games consoles for different family types.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 288322 (Wattalyst). We also would like to the anonymous reviewers for their helpful comments.

References

[1] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[2] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[3] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(0):53 – 65, 1987.

"The *cluster consistency* and the *discriminative index* measures are novel and useful for analyzing such data types that can be adopted and expanded by other data mining practitioners." — Anonymous Reviewer