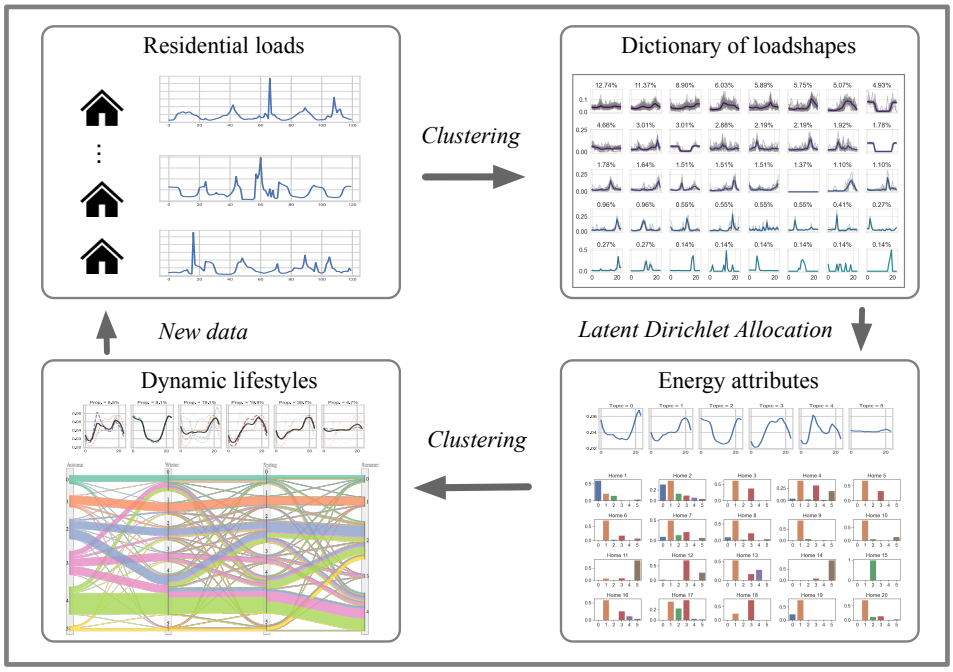
Graphical Abstract

Constructing dynamic residential energy lifestyles using Latent Dirichlet Allocation

Xiao Chen, Chad Zanocco, June Flora, Ram Rajagopal



Highlights

Constructing dynamic residential energy lifestyles using Latent Dirichlet Allocation

Xiao Chen, Chad Zanocco, June Flora, Ram Rajagopal

- We introduce a novel approach for energy lifestyle analysis using Latent Dirichlet Allocation (LDA).
- Using LDA we extract six distinct latent energy attributes from 60,000 households with hourly electricity data.
- We derive six energy lifestyle profiles from energy attributes and describe dominant features.
- We examine seasonal dynamics of energy lifestyles revealing that some households maintain a single lifestyle profile, others multiple lifestyles.
- We describe practical applications including residential energy program design, targeting, and lifestyle change analysis.

Constructing dynamic residential energy lifestyles using Latent Dirichlet Allocation

Xiao Chen[a,1], Chad Zanocco[a,1], June Flora[a], Ram Rajagopal[a,b]

[a] *Civil & Environmental Engineering, Stanford University, Stanford, 94305, CA, U.S.*
[b] *Electrical Engineering, Stanford University, Stanford, 94305, CA, U.S.*

Abstract

The rapid expansion of Advanced Meter Infrastructure (AMI) has dramatically altered the energy information landscape. However, our ability to use this information to generate actionable insights about residential electricity demand remains limited. In this research, we propose a new framework for understanding residential electricity demand by using a dynamic energy lifestyles approach that is iterative and highly extensible. To obtain energy lifestyles, we develop a novel approach that applies Latent Dirichlet Allocation (LDA), a method commonly used for inferring the latent topical structure of text data, to extract a series of latent household energy attributes. By doing so, we provide a new perspective on household electricity consumption where each household is characterized by a mixture of energy attributes that form the building blocks for identifying a sparse collection of energy lifestyles. We test this approach by running experiments on one year of hourly smart meter data from 60,000 households and we extract six energy attributes that describe general daily use patterns. We then use

Email addresses: markcx@stanford.edu (Xiao Chen), czanocco@stanford.edu (Chad Zanocco), jflora@stanford.edu (June Flora), ramr@stanford.edu (Ram Rajagopal)

[1]Equal Contributors

clustering techniques to derive six distinct energy lifestyle profiles from energy attribute proportions. Our lifestyle approach is also flexible to varying time interval lengths, and we provide an example of our lifestyle approach applied seasonally (Autumn, Winter, Spring, and Summer) to track energy lifestyle dynamics within and across households. These energy lifestyles are then compared to different energy use characteristics, and we discuss their practical applications for energy program design and lifestyle change analysis.

Keywords: Energy Lifestyles, Residential Electricity Use, Smart Meter, Latent Dirichlet Allocation, Topic Modeling, Clustering

PACS: 0000, 1111

2000 MSC: 0000, 1111

---

1. Introduction

The growth of advanced metering infrastructure (AMI) has greatly expanded our potential to analyze household electricity usage. To date, AMI infrastructure provides hourly and sub-hourly electricity usage patterns via smart meter technology for tens of millions of households in the United States alone, with the deployment of residential smart meters increasing yearly by millions [1]. Prior analysis of smart meter data has provided insights about household daily load shapes and the variation of electricity use patterns both within and across households. While such information about household energy use patterns is being applied toward forecasting residential demand [2, 3, 4], other nascent applications of smart meter data are of increasing interest to energy providers, including targeting households for demand response (DR) [5, 6, 7], tailoring information to differing user segments about energy efficiency (EE) programs [8], and making recommendations to

customers for enrolling them into pricing programs, such as time–of–use rate plans [9, 10]. Moreover, in many extant applications, household energy use pattern information is typically treated as static, without consideration into how patterns may change across time, either cyclically (e.g., seasons, school calendar year, etc.) or as structural household shifts (e.g., new household members, change in work hours, etc.), potentially missing opportunities for more refined targeting, tailoring, and other program design considerations.

As smart meters become more ubiquitous in households across the world, new advances are needed to process the deluge of data streams produced from these devices and then generate actionable insights—especially in a way that has small computational overhead and does not require continuous human–in–the loop interactions [11]. In particular, using smart meter data to understand household level energy use is an on-going challenge, and with it comes the difficulties in developing meaningful interventions that can ease burdens on the grid while also contributing to energy system decarbonization and maintaining customer engagement and satisfaction [12]. For example, one such motivation for household–level energy interventions is to reduce greenhouse gas emissions from nonrenewable generation sources (e.g., natural gas "peaker" plants) during periods of high demand, while also expanding the potential for customers to change their energy behaviors and appliance purchases to save money on their energy bills [13].

While existing work on customer identification and segmentation has been explored in the literature [6, 14, 5, 10], insights about customer segmentation for households and groups of households do not demonstrate strong linkages to a variety of common occupant behaviors with few exceptions [15] (see section 2 for examples). Additionally, while many methods have been proposed for gaining broad insights into customers and groups of customers'

electricity consumption, these methods are often too complex to easily scale for use by utility companies or energy providers [16]. Existing methods usually require higher-resolution data than what is typically available via smart meters, and these methods have not produced the sort of broadly generalizable insights needed to effectively inform program design [17].

To address these challenges, we use Latent Dirichlet Allocation (LDA) to analyze daily energy consumption patterns of households. While LDA is most commonly associated with Natural Language Processing tasks such as extracting latent topics from text-based documents [18, 19], it is now increasingly applied in other domains and problem areas unrelated to text analysis, including genomics and the remote sensing of satellite imagery [20, 21].

We propose yet another application for LDA previously unexplored: the classification and interpretation of energy use patterns in the home. In doing so, we seek to identify latent patterns of daily energy consumption and then use these latent constructs to build residential energy lifestyle profiles. Our approach does not directly characterize residential energy activities and behaviors through observational or self-reported methods [22, 23] or real time data disaggregation of household energy consumption–all of which can introduce complexity that makes it challenging to generalize across households. Instead, our conceptualization of energy lifestyles are more broadly construed, with the potential to generate meaningful insights without having to resort to finer grained, more nuanced understandings of energy use and energy–related activities within the home. Such an understanding of energy lifestyles could have applications for energy practitioners, such as electricity service providers, policymakers, and the research community for tasks including identifying energy use patterns, targeting customers, and under-

standing household demand flexibility and response of residential users.

Our approach toward developing energy lifestyles also affords us new opportunities in examining the temporal dimensions of lifestyles, or how these energy lifestyles may change across time intervals of varying length. Previous research has considered a lifestyle as a static attribute of a household, with the lifestyle referring to a component that does not change across time. However, research suggests that lifestyles can indeed have dynamic components [24, 25, 26, 27, 28, 29], yet much of this literature is limited to within–day time organization as opposed to across days, weeks, months etc. On a global scale, we have recently experienced dramatic disruptive influences that has changed the nature, organization, and amount of electricity consumed–lifestyle changes that have occurred during the COVID-19 pandemic [30, 31, 32]. While the measurement of lifestyle change through electricity use may only serve as an approximation for a variety of conditions and activity patterns that occur within a household across time, we postulate that such a lifestyle approach could provide a signal for when large changes related to energy use occurs in the home. Such changes could include anything from a change in the number of household occupants (e.g., a child being born or leaving for college) to a change in the patterns of occupancy (e.g., new employment or retirement) to broader "shocks" such as recent COVID-19 related restrictions. On the other hand, some households may experience little to no change in energy lifestyles across time, also providing an important insight into the stability of energy use patterns and their associated household activities. Understanding these characteristics of lifestyles could bring new opportunities for energy providers to dynamically target energy programs during certain times throughout the year, and allow the iterative identification of lifestyle patterns based on constantly updating

data streams from AMI infrastructure. While this understanding has the potential to both improve recruitment and engagement in efficiency and demand response programs [33, 7], we may also find that this more "real" life understanding of residential consumption leads to the development of new policies and programs.

In this research we break new ground in constructing temporally dynamic energy lifestyles using a novel application of LDA. Given this focus on generating and gaining insights from temporally dynamic energy lifestyles, our research seeks to answer the following research questions: (1) What residential energy lifestyle profiles emerge from empirical data and what are their prominent characteristics? and (2) What patterns of change, or stability, is observed in households' lifestyle profiles across time and what is related to these temporal dynamics?

We address these research questions in the following sections. First, we describe our approach for generating energy lifestyles by introducing our conceptual framework and method. Next, we describe our residential electricity dataset and experimental setting. Then we derive energy lifestyles and provide insights about their prominent features and patterns of change across time. Lastly, we discuss applications of this lifestyle approach and provide recommendations for future research.

2. Constructing Energy Lifestyles

2.1. Conceptual framework

While there are many ways to describe a lifestyle, we adopt the definition that a lifestyle is the consolidation of a persistent set of patterns of behavior that occur within the home environment [34]. We propose that energy

consumption is best understood as a consequence of lifestyles that reflect the organization, sequencing, synchronicity, habitualness, and contingent or interdependence of the timing of the activities of daily life within a day and over weeks, months and years.

To capture the temporal patterns in energy use, our conceptual approach to energy lifestyles is built around the daily load shape as a core feature of household consumptive patterns, which imparts information about the relative magnitude, duration, and timing of energy use throughout a day (24 hour period). Embedded within this daily load shape representation is information about energy use related to the timing of household activities (e.g., cooking, cleaning, entertainment), appliance characteristics (e.g., heating/cooling technologies), household characteristics (e.g., number of occupants, age of occupants, etc.) and contextual and environmental characteristics (e.g., weather, climate). Features of the daily load shape, including the timing of peak, base, and the ratio of peak/base, contribute to insights about the relation between activities and electricity use. Finally, the variation of load shape patterns (i.e., entropy) imparts information about consistency or inconsistency of energy use patterns across time. The load shape itself, therefore, contains rich information about a household's energy use and everything within the household related to this use.

To encompass this broad representation of household energy use with daily load shape as a focus, we envision a framework that applies Latent Dirichlet Allocation (LDA). LDA is a generative statistical model that allows unobserved groups to be explained by a set of observations that have related characteristics. The canonical example of LDA is identifying topics in text analysis [35, 36, 37], where words are observations that are collected from documents, such as a newspaper article, and each document is some

mixture of topics that can later be assigned semantic meaning (e.g., politics, sports, etc.). In the text-based example, the process of generating a document is described by a sampling of topics from a mixture of topics, and a sampling of corresponding words according to those topics, and then repeating this process to generate all words in the document. Topics are initialized randomly and then updated through iterations using Variational Bayesian Inference [18, 38] or Markov Chain Monte Carlo [39] approaches until a convergence criteria is met, where convergence is measured by the change of likelihood in producing the observations (words), or the change of the inferred parameters. This topic modeling approach has now been applied in many fields including environmental science for understanding policy change in dealing with climate change [40] and biology for tasks ranging from extracting common patterns of mass fragments to neutral losses in molecules [37, 41].

Table 1: Conceptual relationship between text analysis and energy analysis in a LDA setting

Text and language		Residential energy consumption
documents	$\longleftrightarrow$	household
words	$\longleftrightarrow$	load shapes
topics	$\longleftrightarrow$	energy attributes

Applying LDA to the context of analyzing energy demand, we develop a novel application that extracts latent patterns of energy consumption by considering households as documents, and treating load shapes as words. A conceptual relationship of terms in the domain of language analysis and in the case of our proposed energy analysis is shown in Table 1. In this

comparison example, latent energy patterns, which we have named an energy attribute, is synonymous with a topic in the text analysis example.

This framework is expected to generate two nested components. The first component is the aforementioned energy attribute, a latent characterization of daily energy use patterns. These energy attributes are derived from daily load shapes and form the building blocks of dominant daily energy use patterns in a household, and each household can contain different proportions of these latent attributes. When proportions of these energy attributes are aggregated across a large pool of households, their mixtures among certain household-types can be used to generate the second layer of abstraction which we refer to as an "energy lifestyle". Energy lifestyles, therefore, are an expression of collective daily energy use patterns across groups of households. These energy attributes and energy lifestyles can be applied to any temporally consumptive datastream (e.g., electricity and gas) dependent on the availability of such information.

2.2. Overview of methods

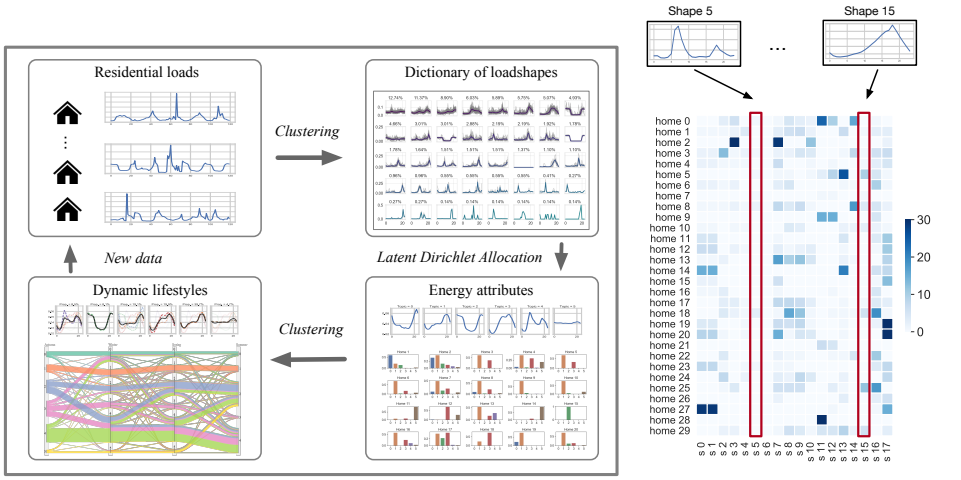
In the context of analyzing energy demand, we develop a method to extract latent patterns using LDA. In our energy analysis case, analogous to the document example where each document contains a mixture of topics, we assume that each household contains a mixture of energy attributes. Therefore, an objective is to identify latent attributes of energy consumption across many households and construct load shapes denoted as s. Specifically, for a j-th home having a mixture of K attributes, the household's attribute mixture weights θ_j is a probability distribution drawn from a Dirichlet prior with parameter α and the k-th attribute is a multinomial distribution ψ_k over a S-shape vocabulary (or dictionary). For i-th shape s_{ji} in home j, a

topic $z_{ji} = k$ is sampled from θ_j and s_{ji} is drawn from ψ_k. The generative model can therefore be expressed as

$$\theta_j \sim Dir(\alpha), \psi_k \sim Dir(\beta), \{z_{ji} = k\} \sim \theta_j, s_{ji} \sim \psi_k \quad . \tag{1}$$

We briefly describe the LDA method here to build intuition about its application and then we expand upon this by providing a more detailed description in Appendix A.8. Once energy attributes are finalized, we then apply the k-means clustering method on the energy attribute space of households in the second stage to generate a sparse representation of energy usage patterns over days (characterized by cluster centers), which we refer to as energy lifestyles because they contain latent patterns of energy usage generated across households. To provide an overview of the entire process of generating energy lifestyles, we constructed a simplified workflow displayed in Figure 1a, described in detail below.

To dynamically create and analyze energy lifestyles, we gather a collection of daily load shapes that covers a majority of patterns of household consumption by clustering over raw meter data (Figure 1a). Such a collection forms a load shape dictionary that allows us to identify the frequency of generalized load shapes for each household (Figure 1b). After obtaining the frequency counts of shapes for households, we use the LDA method to yield representative latent energy attributes. Correspondingly, the households can also be viewed as mixtures of attributes (Figure 1a). Because many households share similar energy attribute distributions, we use a clustering method to group similar households in their energy attribute space, yielding distinct clusters (i.e., lifestyles). Each cluster represents an energy lifestyle group that can be further interpreted using the proportions of attributes of energy use patterns within the cluster. In addition, to explore temporal



(a) Schematic flow of generating dynamic lifestyles.

(b) Frequency counts of a few shapes

Figure 1: (a) The process of generating dynamic lifestyles. We first cluster raw smart meter data to create a dictionary of load shapes. Next, we apply LDA to identify representative energy attributes. Based on these attributes, we finally compose the lifestyles across time. (b) We present the frequency counts of a subset of load shapes (denoted as s 0 to s 17 on x-axis) for some sampled homes (*home* 0 to *home* 29 on y-axis) over the Autumn season (Sept. - Nov.).

dynamics we apply the LDA method on seasonal intervals (Autumn, Winter, Spring, Summer) and assign each household to its nearest computed lifestyles, which we display as seasonal transitions of lifestyles for households (Figure 1a). We can then iterate on these steps and re-generate insights of household consumption patterns as data streams are updated. While in this research we explore temporal dynamics seasonally, this method can be applied across other time intervals (e.g., monthly, bi-annually, annually, etc.).

3. Experiments and Results

Our framework is heavily driven by empirical data from actual residential households using several contemporary machine learning methods. Specifically, we first run an experiment for generating a representative load shape dictionary by clustering raw residential smart meter data. In our next series of experiments, we apply this load shape dictionary and then synthesize typical lifestyles by using LDA. When summarizing the lifestyle profiles, we assign them names according to a composite shape formed via reconstructing the weighted sum of load shapes. Finally, we run set of experiments to validate the profiles by examining the electricity consumption features and show these features (e.g., the ratio of morning to whole day energy use) support temporal characteristics of these lifestyles. We describe our data and approach for analyzing this data in detail in the following sections.

3.1. Description of residential smart meter data

Our work utilizes a large dataset of residential load consumption from Pacific Gas & Electricity (PG&E) spanning from August 1, 2010 to August 1, 2011, which contains more than one hundred thousand customers. For analysis, we randomly selected 60,000 homes having hourly smart meter

data, comprising 436 ZIP codes in California, U.S.A, covering eight different climate zones (Figure A.12). Such a sample population, which is larger than many previous studies [42, 43], is appropriate for capturing heterogeneity in residential energy consumption patterns. From this data, we then convert each household's load shape pattern into the format of (24-hour) daily sequences over the course of a year for our analysis.

3.2. Dictionary of load shapes

Since households display a variety of load shapes across time, and that the mixture of these load shapes is associated with the lifestyle that households may possess, we first learn a dictionary of daily load shapes that is the foundation of our energy lifestyle approach. To generate a robust dictionary of load shapes, we utilize clustering methods with a careful selection of distance metrics (Appendix A.2.2).

Given a set $\mathcal{X}$ that includes all daily electricity loads and a data point $\boldsymbol{x} \in \mathcal{X}$, we would like to find a number of representative points of clusters, denoted as a set $C \supseteq \mathcal{X}$, that can summarize a massive dataset into several typical patterns. To accomplish this, we minimize the distance between points and cluster sets $\min_{\boldsymbol{x}\in\mathcal{X}} d(\boldsymbol{x}, C)$ in metric d, where $d(\boldsymbol{x}, C) = \min_{c\in C} d(\boldsymbol{x}, c)$ is the minimum distance from $\boldsymbol{x}$ to a center c. Taking the standard k-means as an example, we have an assignment $\phi : \mathcal{X} \to C$ of points to clusters so as to $\min_{\boldsymbol{x}\in C} d(\boldsymbol{x}, \phi(\boldsymbol{x}, C)) = \min_C \sum_{i=1}^{k} \min_{\boldsymbol{x}\in C_i} ||\boldsymbol{x} - c_i||_2^2$, where d is the Euclidean distance between two points. Besides the Euclidean distance, we also apply the cosine distance, the L^1 distance, and the dynamic time warping (DTW) distances to perform the clustering for the load shape dictionary. We also test k-medians [44], hierarchical clustering [45], and DBSCAN [46] clustering methods for comparison (see Appendix A.2.1 for more informa-

tion). We set the load shape dictionary of size 200 using the k-medians method with a hybrid of DTW and Euclidean distances, because this setting yields a good coverage of profiled shapes with the highest score on the Calinski-Harabaz Index [47]. Further explanation is provided in Appendix A.2

3.3. Energy lifestyles composition

Once we have derived a dictionary of 200 load shapes, we use the clustered labels (i.e. shape indices) to represent each household's load shape pattern. Specifically, we calculate the frequency of the load shapes for a household and represent them as a 200-dimensional vector. For example, during a calendar year, if the home i repeated shape1 for 23 days, shape2 for 17 days, shape200 for 325 days, then we have the vector $r_i = [23, 17, \ldots, 325]$ to describe the one-year pattern of home i, where $r_i \in \mathbb{R}_+^{200}$. Referring to Figure 1b, we stack all households' load patterns into an n-by-200 matrix M_r where n is the number of homes.

We apply the LDA method to extract a few distinct and representative attributes of load shapes. To determine how many attributes are appropriate to both capture all consumption patterns while also being sufficiently representative, we prescribed 10 attributes and then merge the neighboring attributes together using a bottom-up approach, i.e. by calculating correlations of attributes and projecting them down to lower dimensions. After consolidating similar attributes, this ultimately yields six representative attributes that are quantitatively and descriptively distinct in terms of daily consumption patterns (Figure 2). More details including why we chose 6 attributes are covered in section Appendix A.3. We observe that *attribute 0* has the peak consumption around 10pm-11pm with low energy use dur-

ing the day. A similar pattern is also observed in *attribute 2* but with a longer time span of late-night consumption, whereas *attribute 1* has a peak consumption around 6pm-7pm with lowest energy use around 2am-3am. Distinct from other attributes, *attribute 3* has the highest energy usage in the afternoon from 12pm-5pm and *attribute 4* displays peak usage around 7am-8am in the morning. Finally, we find *attribute 5* has comparatively low variation in daily usage.

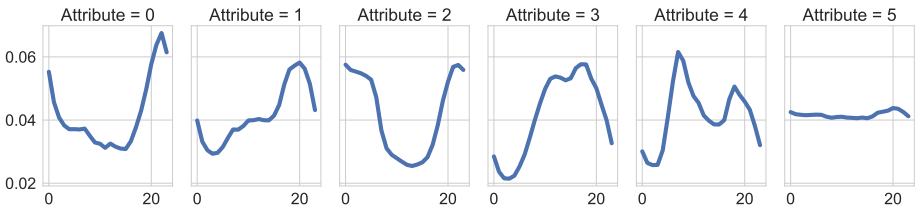


Figure 2: Attribute shapes. Each attribute shape is a weighted sum of 200 dictionary shapes, where the weights are the normalized probabilities of each shape's occurrence.

With these six summarized attributes, each home is then characterized by assigning a 6-dimensional vector where the value at each entry represents the corresponding attribute weight. The attribute weight at the k-th entry indicates how likely a home possesses *attribute-k*. In our experiment, many households share similar attribute weights over a year, so we therefore define lifestyles using k-means clustering to obtain a stable result of six lifestyles. We found that six lifestyles were sufficient to cover the heterogeneity of daily lifestyle patterns (shown in Figure A.16 and Figure A.17). In Figure 3, we plot each lifestyle as a black line that represents a weighted average of different attribute weights depicted by the thickness of the dashed lines. Given their load shape characteristics, for ease of reference we assigned names to each of the lifestyles from left to right as *active mornings, night owl, ev-*

eryday is a new day, home early, home for dinner, and *steady going*. In naming these lifestyles, we use the following as descriptive justification: *active morning* has a distinguishing characteristic of energy use in the morning time period; *night owl* has energy use in the very late night and very early morning with little morning through evening usage across days; *everyday is a new day* displays substantial heterogeneity in energy use across the day; *home early* is distinguished by its late afternoon use; *home for dinner* has energy use concentrated in the evening; and the *steady going* lifestyle has use that remains relatively stable throughout the day. We have no additional, non electricity-use information about these households to verify or justify these lifestyle names, a challenge confronted by other "unsupervised" learning applications [37, 43].

The *home for dinner* lifestyle is the most frequently occurring lifestyle among our sampled residential households, accounting for nearly 40% of the households in our dataset. The next most frequently occurring are the *home early* and the *everyday is a new day* lifestyles, which account for 19.1% and 19.9% respectively, followed by *active mornings* and *night owl*, both of which account for approximately 8% of the sample. At 4.7%, the least frequently occurring lifestyle is *steady going*. Although the mean representations of *everyday is a new day* and *home for dinner* are similar, they are different in that the mixture weights of the attributes are evenly distributed for *everyday is a new day* but the weights for *home for dinner* are concentrated on *attribute 1* (Figure A.17).

While different households have different lifestyles, we also observe that a single household can also display multiple lifestyles over the course of a year. For example, one set of lifestyle patterns could be related to the presence of children in the home, such as when children are on break during the

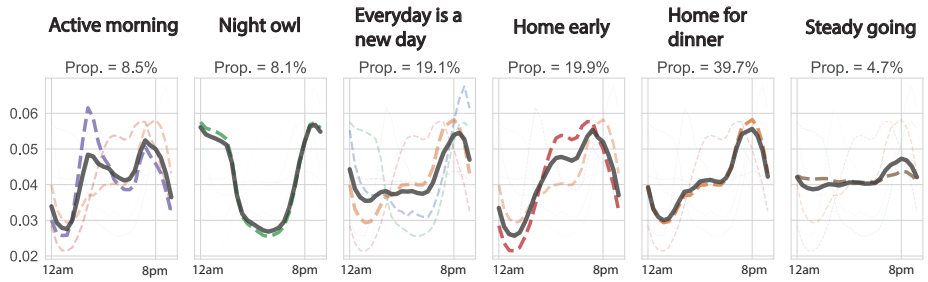


Figure 3: Lifestyles. From left to right, they are *active morning*, *night owl*, *everyday is a new day*, *home early*, *home for dinner*, and *steady going*. The different thickness of the dashed lines indicate different composition weights for corresponding attributes.

summer months and in school during the fall. The change of lifestyles of a household relates to its energy use that could be associated with a household members' behaviors (e.g., occupancy) under different time horizons (e.g., months, seasons, years, etc.). To this end, we next examine how these energy use behaviors change across time by choosing season as a convenient unit of measurement, as we only have access to one year of data and cannot observe changes over longer time periods. Therefore, we partition our one year's worth of data into four seasons: Autumn (Sept. - Nov.), Winter (Dec. - Feb.), Spring (Mar. - May), and Summer (June - Aug.) and run lifestyle analysis for seasonal time intervals. The lifestyle transitions of households across seasons is displayed in Figure 4.

We find that the *home for dinner* comprises a larger proportion of household across seasons compared to the other lifestyles. Such a seasonal phenomenon also matches the previous findings in the observations across the entire year (in Figure 3). Each season contains households with all six lifestyles except for summer which does not contain any households with the *steady going* lifestyle. One reason could be that the relatively flatter us-

age profile of *steady going* is particularly uncommon during summer months because thermal comfort-related consumption—such as HVAC usage—tend to be turned on and off for a multiple hours across a day, yielding a more volatile daily load shape. Whereas in the winter, many homes rely on gas–heating and regulation of thermal comfort could yield a flatter pattern of electricity use. A detailed description of household sample membership in lifestyles across the four seasons is provided in Table A.4. Furthermore, we find that some households stay in a single lifestyle across all seasons, whereas other homes switch between two or more lifestyles over the course of four seasons. Such an observation motivates us to investigate the distinctions between those lifestyle-consistent households and lifestyle-changing households.

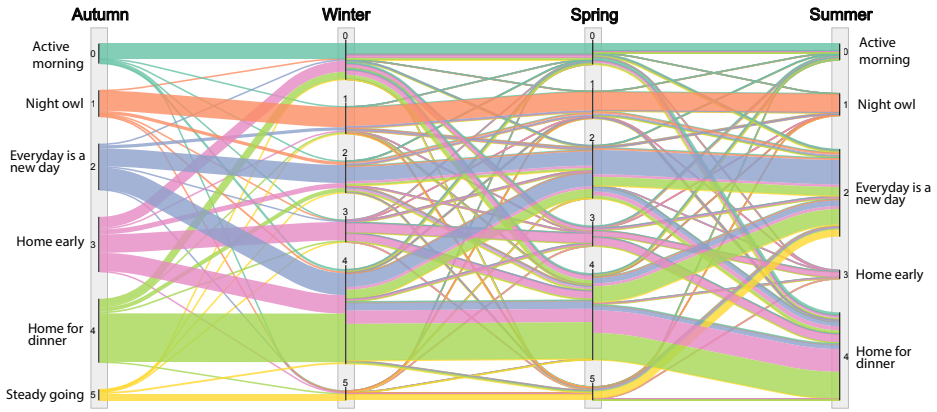


Figure 4: Seasonal transition of lifestyles spanning from Autumn 2010 through Summer 2011. The color of the lines represent the lifestyle designation in Autumn and tracks groups of households across time. The thickness of the lines represent the proportion of total households in each lifestyle at each seasonal interval, with wider lines indicating more households and thinner lines fewer households.

3.4. Energy lifestyle analyses

To understand what determines different lifestyles, we explore a number of energy use characteristics derived from raw smart meter data from households. Unlike many other studies [15, 43, 48], our energy use characteristics are generated using raw energy data from households. These energy use characteristics (also known as features) are directly derived from raw smart meter data, which does not rely on the previously generated load shape dictionary or energy attributes. We first illustrate the features associated with corresponding lifestyles summarized across one year. Next, we explore the changes overtime of various features of lifestyles across seasons, and develop a way to identify those households that change lifestyles across seasons (*Changer*) and those who do not (*No Changer*).

3.4.1. Features of energy use

Once we have constructed our lifestyles, each home is automatically associated with a single lifestyle that matches its energy use pattern. Conditioning on these lifestyles, we consider all households that are labeled in the same lifestyle as a group, yielding six groups among all households in our sample. We find that each group of households has unique distributions of certain energy use features. These features are extracted from raw electricity use from households, such as mean daily energy use, ratio of morning to whole day energy use, and peak hour frequency (normalized), described in detail in Table 2. We are interested in these distributions of features, especially when they have distinct characteristics, since they can be used to identify different lifestyles from the pattern of features instead of observing a whole year of household consumption.

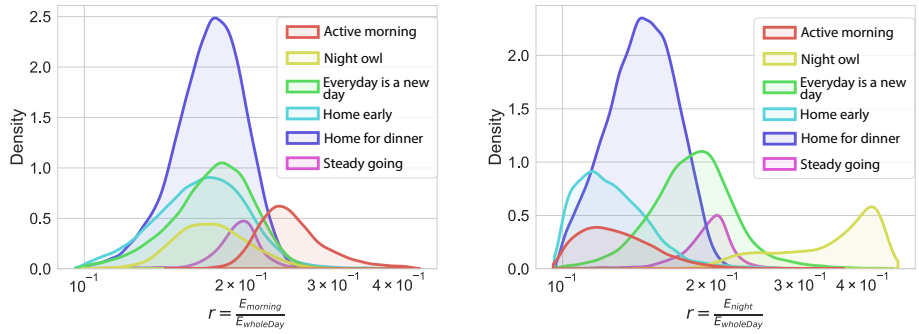
We present a few examples showing that certain lifestyles can be distin-

Table 2: Description of features of electricity use

Feature	Description
E_{day}	mean of daily energy use
E_{hour}	mean of hourly energy use
E_{peak}	mean energy use of peak hour in a day, equivalent to E_{max}
E_{base}	mean base energy use of a day
E_{min}	mean of min energy use of a day
$E_{morning}$	morning energy use between 6am to 10am
E_{noon}	morning energy use between 10am to 2pm
$E_{evening}$	evening energy use between 6pm to 10pm
E_{night}	night energy use between 10pm to 2am
$E_{wholeday}$	24 hour energy use
r_{base}	base load ratio, i.e. mean of $\frac{E_{base}}{E_{day}}$
$r_{min2max}$	mean ratio of min hourly load divided by max hourly load, i.e. mean of $\frac{E_{min}}{E_{max}}$
r_{m2w}	mean of morning energy use divided by whole day energy use, i.e. mean of $\frac{E_{morning}}{E_{wholeday}}$
r_{n2w}	mean of noon energy use divided by whole day energy use, i.e. mean of $\frac{E_{noon}}{E_{wholeday}}$
r_{e2w}	mean of evening energy use divided by whole day energy use, i.e. mean of $\frac{E_{evening}}{E_{wholeday}}$
r_{ni2w}	mean of night energy use divided by whole day energy use, i.e. mean of $\frac{E_{night}}{E_{wholeday}}$
π_j	multinomial distribution over 24 hours showing the normalized frequency of peak hour occurrence. The j takes value from 0, 1, ..., 23, indicating j-th peak hour in a day

guished from feature distributions (e.g. Figure 5a and Figure 5b). Specifically, Figure 5a indicates that the *active morning* group has the highest ratio of morning (6am - 10am) to whole day energy use where the mean is approximately 0.24 and with a small portion of the homes having a ratio value over 0.4. The *home for dinner* group has a mean ratio value of approximately 0.18 with a substantial portion of homes having an even lower value of 0.1. Other lifestyles have the mean ratio value between 0.16 - 0.19. In short, these distributions are consistent with initial insights about our lifestyles as *active morning* has higher energy use than other lifestyles in the morning time period. The *night owl* style has the highest ratio of night (10pm-2am) to whole day energy use where the mean is approximately 0.44, suggesting that many homes use energy between 10pm-2am, accounting for approximately 44% of the whole day use in that lifestyle group. Other styles have mean ratio values below 0.15 except for *everyday is a new day* and *steady going* lifestyles, both of which have either non-trivial energy consumption during the night period because of switching between different energy attributes or have a flatter shape associated with the energy attributes.

Apart from the intra-day's ratio of energy use, we compare the peak hour occurrence of the different lifestyles. We present four typical lifestyles (Figure 6) because they are representative and prevalent among households. Figure 6 suggests that the distributions of peak hour frequencies align with lifestyles even though the frequency of occurrences are extracted from raw energy use. For example, the pattern of peak hours frequency for *night owl* (Figure 6b) closely matches with its lifestyle curve (Figure 3). Similar matches can also be found in other lifestyles like *active morning*, *everyday is a new day*, and *home for dinner*. Such descriptive cross-validation in peak hours demonstrates the value of our lifestyle framework, while also



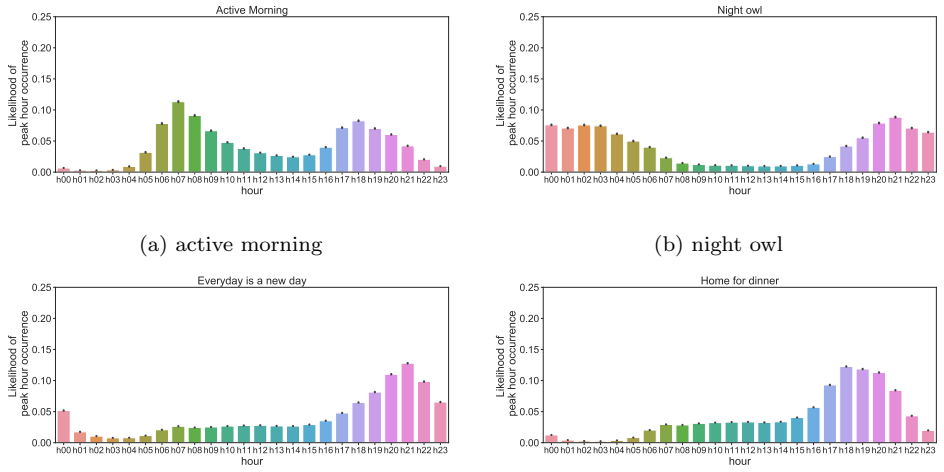
(a) ratio of morning to whole day energy use

(b) ratio of night to whole day energy use

Figure 5: Load features of different lifestyles. (a) suggests that *active morning* style has a higher ratio of morning to whole day energy than other lifestyles. (b) reflects that *night owl* has a significantly higher ratio of night to whole day energy than other lifestyles. One potential benefit of looking into these ratios is that we can quickly classify some homes into corresponding lifestyles without observing their annual consumption data.

building an understanding around inductive features for various lifestyles. We present additional summaries of features in Appendix A.5.

As the distributions of features differ substantially among various lifestyles, we expect that lifestyles can be identified by using load-related features. To assess this, we establish a classification problem where the lifestyle of household i is the label y_i and the features are the observed predictors $\boldsymbol{x}_i$. We therefore learn a mapping f such that $y_i = f(\boldsymbol{x}_i), \forall i \in 1 \dots N$ where N is the number of samples. For interpretability and robustness, we apply random forest (RF) as our classification model. After splitting the training, validation, and test sets using the portions of 70%, 10%, and 20%, followed by selecting and calibrating features, we then fit a RF model with a classification accuracy of 68.5% on average (in Figure 7a). We find that *night owl* is the easiest lifestyle to classify having approximately 82% accuracy. In contrast, *home early* is the most difficult lifestyle to model with 47% accuracy



(a) active morning

(b) night owl

(c) everyday is a new day

(d) home for dinner

Figure 6: Peak hour distribution. Each sub-figure shows the averaged frequency of peak hour occurrence for all homes in the corresponding style group. We illustrate four lifestyles because these lifestyles are prevalent among the households.

since a significant portion is miss-classified as *home for dinner*. These observations are also supported by the classification results of precision, recall, and F1 score (shown in Table A.5).

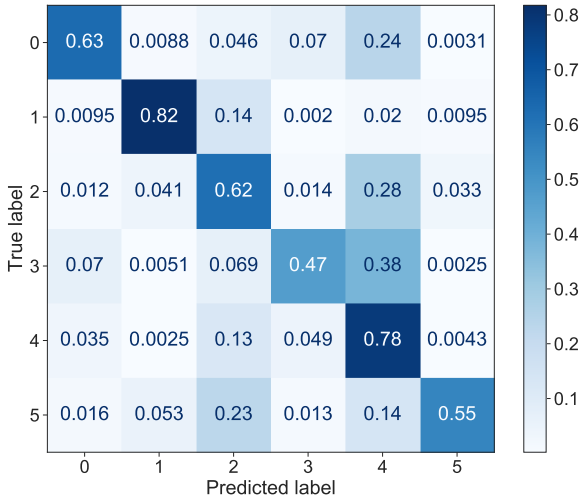
In addition to comparing the feature distributions of lifestyles and classifying each lifestyle based on household energy consumption, we investigate what features have important roles in determining lifestyles. We use a model-agnostic permutation importance score described in [49, 50] to estimate the importance of the features in our random forest model, and discover that the features constructed as various ratios play major roles in identifying lifestyles (Figure 7b).

We find the mean ratio of night to whole day usage is the most important feature, contributing to approximately 18% of additional accuracy compared to a case where the ratio is identically distributed (i.e. random assignment),

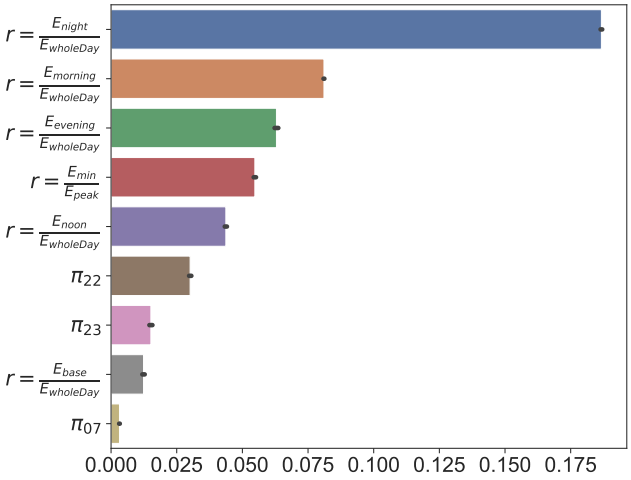
followed by the mean ratio of morning to whole day that contributes an additional 8% accuracy. We also observe that the peak hour frequencies at 7th, 22-th, 23-th hour are non-trivial in determining the lifestyle, suggesting that the peak consumption in the night around 10pm-11pm and in the morning around 7am are important features. As an additional robustness step, we verify that these top features are not highly correlated (see Figure A.27). We further verify these results by running multinomial logistic regression models and find statistical significance for these ratio features. We provide more details in section Appendix A.7.

3.4.2. Dynamics in energy lifestyles across time

In addition to these year-specific patterns, we also compare the distributions of features at the seasonal-level because certain households may change lifestyles (i.e., Changer) or may not change lifestyles (i.e., No Changer) across a single year period. Since the *steady going* lifestyle does not occur in the summer (Figure 4), the following analysis is focused on the remaining lifestyles (Figure 3). First, we assess the characteristics of No Changer households in terms of load-related features. In particular, we compare both the ratio of morning to whole day usage and the ratio of evening to whole day usage across four seasons to check the stability of the feature distribution among various lifestyle groups. We find that *active morning*, *everyday is a new day*, *home early*, and *home for dinner* have very stable distributions across four seasons. Consistent with the lifestyle name, *active morning* is influenced by the morning to whole day ratio value (approximately 0.26) compared with any other lifestyle's mean ratio (that is below 0.2) (Figure 8). Although *night owl* households tend to keep this lifestyle across multiple seasons, we note that the ratio of morning to whole day usage of the *night owl*



(a) Confusion matrix of classifying lifestyles



(b) Feature importance

Figure 7: Classification results. (a) The confusion matrix suggests that *night owl* has the highest accuracy at 0.82. In contrast, the *home early* has the lowest accuracy, 0.47, since a majority of samples are misclassified as *home for dinner*. (b) indicates the top nine important features needed to correctly classify a home's lifestyle. In this case, the ratio of night to whole day energy use is the most important feature.

lifestyle shifts toward smaller values in the summer compared to other seasons, indicating some homes either increase their whole day energy use or reduce their consumption in the morning period during the summer. Such a pattern matches with previous discoveries [51] that large consumption or late morning activities are more likely to occur in summer for residential households. To confirm the No Changers' stability of load characteristics, we further compared the ratio of evening to whole day usage across the seasons in Figure 9. We observe that all lifestyles have stable distributions of this ratio, with means located between 0.27 to 0.32, except for the *night owl* style that has a mean of 0.19 in summer and 0.34 in winter. Some homes in *night owl* lifestyle indeed show larger consumption after midnight in the summer, whereas in the winter the night usage pattern begins earlier in the night. Other features, such as mean load and peak load, also demonstrate the stability of No Changer households in various lifestyles (in Figure A.24 and Figure A.23).

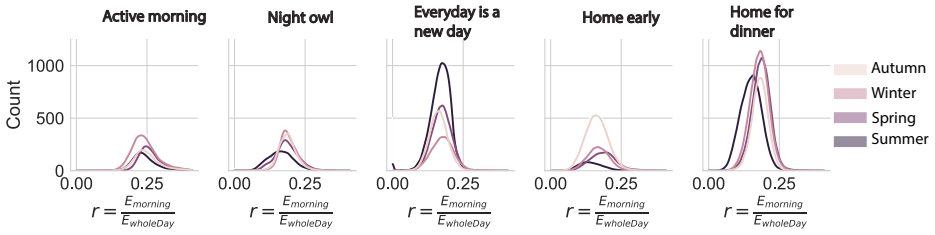


Figure 8: Ratio of morning to whole day of energy. The distribution mode of this feature is relatively stable for No Changer, except during summer season a small portion of population shifts the ratio a little in different lifestyles.

Second, we compare the distributions of load features between Changer and No Changer to understand the difference between these two groups across various lifestyles. Specifically, we evaluate these two groups given a

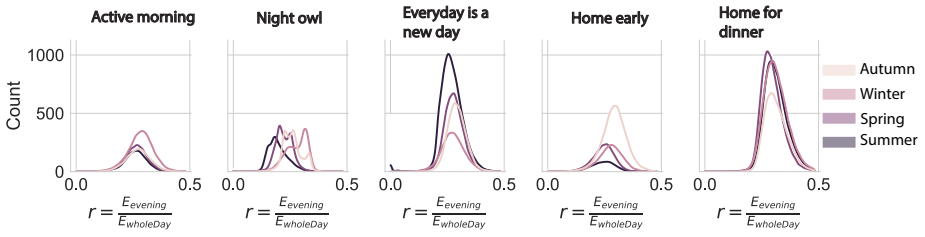


Figure 9: Ratio of evening to whole day of energy. The distribution mode of this feature is relatively stable for No Changer in many lifestyles. Yet in the Night owl lifestyle, some homes still change the usage across four seasons indicated by the second panel above.

lifestyle and a season, and then expand the evaluation over multiple seasons and lifestyles. For example, the distribution of the ratio of morning to whole day usage is expressed in Figure 10, which suggests three insights. First, in the *active morning* lifestyle, the Changers' mean is lower than the No Changers' mean over four seasons. Such a pattern indicates that No Changers tend to consume more in the morning than Changers. Second, overall the No Changers have lower means than Changers for the *night owl*, *everyday is a new day*, *home early*, and *home for dinner* lifestyles in four seasons. When comparing the composition of these attributes, No Changers in those lifestyles have higher consumption in the afternoon than in the morning, indicating that morning usage is relatively small. Thus, the Changers could have higher morning usage because they are not restricted to a single lifestyle. Third, the population of Changers is larger than that of No Changers. Many Changers switch their styles between *everyday is a new day*, *home early*, and *home for dinner*. In the winter, Changers are mainly concentrated in the style of *active morning* and *home for dinner*. In contrast, in the summer, Changers are mainly located in *everyday is a new day* and *home for dinner*. Alternative comparisons using the base–to–peak

ratio (in Figure A.25) also suggest that No Changers differ from Changers across seasons.

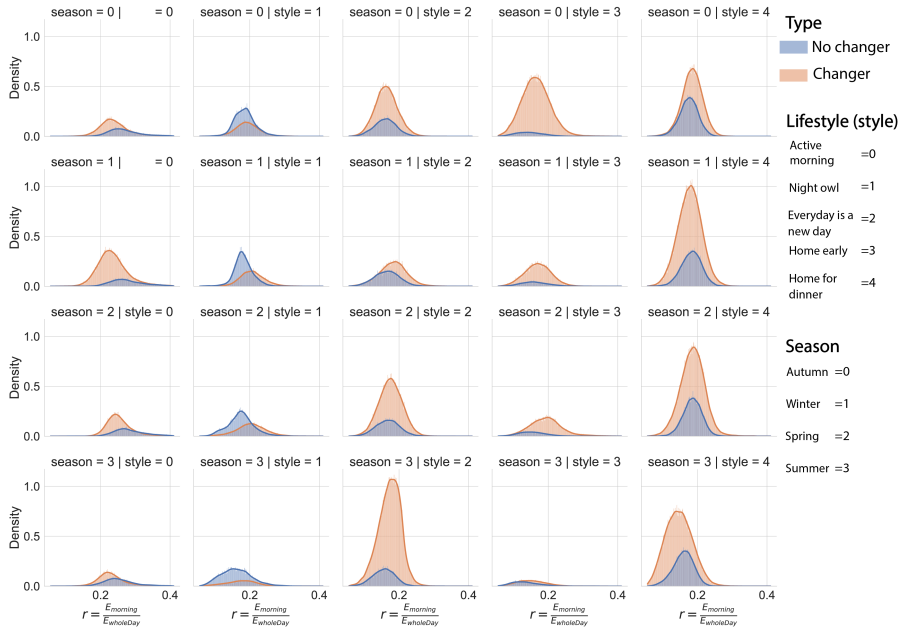


Figure 10: Ratio of morning to whole day of energy. Lifestyle is abbreviated to "style" for visualization purposes.

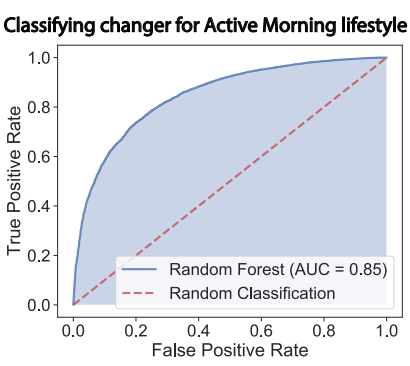
Considering these distinguishing distributions of characteristics between *Changers* and *No Changers*, we then classify these two groups in the context of different lifestyles, because being able to distinguish whether a household is a Changer or No Changer can help energy service providers to identify different types of households and provide customized services (e.g., different energy saving rebates). To accurately classify a Changer or No Changer in each lifestyle, we label the No Changer homes as 0 and label the Changer homes (who possessed the corresponding lifestyle once and then switched to other styles) as 1, and then apply a random forest model to this binary

classification problem. For example, the style of *active morning* achieves 87.9% identification accuracy (Table A.6). Because the positive and negative samples are not evenly distributed, the binary decision can be adjusted for a low false positive rate, as shown in the receiver operating characteristic (ROC) curve in Figure 11a (where the red dotted line denotes performance of random selection). The area under the curve (AUC) is 0.85, meaning that a randomly selected positive example (i.e. a home of a Changer) is more likely to be a Changer than a randomly selected negative example (i.e. a home of a No Changer) with probability 0.85.

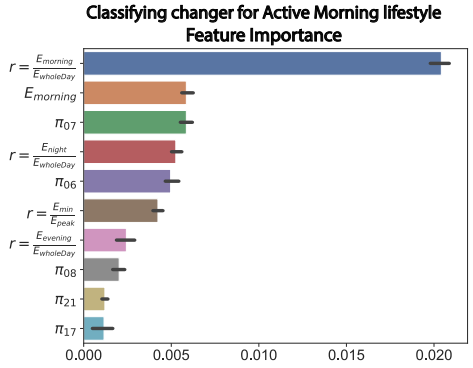
Once the classifier is fitted to identify a Changer, we evaluate the top determinant features by again using the permutation importance method. Figure 11b suggests that the ratio of morning to whole day usage, the morning energy use, and the peak hour frequency at the 7-th hour (7am-8am) are among the top three most important features. Such findings indicate that the pattern of energy consumption in the morning period can largely determine whether a household is a Changer or No Changer in the active morning lifestyle. We also verify that these top importance features are not highly correlated (Figure A.27), which demonstrates the robustness of our results regarding important features.

We assess both the classification performance and feature importance when identifying Changers and No Changers in other lifestyles (Figure A.28, Figure A.29). The results show that classifying Changers in the *night owl* lifestyle has the highest AUC value of 0.97, and doing so in the *home for dinner* lifestyle has the lowest AUC value of 0.77. Such different performances of the AUC metric suggest that identifying Changers in the *night owl* group is much easier than identifying Changers in the *home for dinner* group (Figure A.28a and Figure A.28d). For feature importance, we find

individual lifestyles to have their own prominent features that determine Changers separately, but that the features that are characterized by various ratios of energy use play important roles in all lifestyles (Figure A.29). In general, features related to certain time spans within a day (such as ratio of evening to wholeday energy use) can be applied to identify whether a household is a Changer or not, and have a better performance compared to volume-based features (e.g. base load and hourly mean load, etc.).



(a) By varying the classification threshold, we can trade off between true positive rate (TPR) and false positive rate (FPR). The receiver operating characteristic (ROC) curve shows the TPR and the FPR are significantly higher than random classification, having an AUC of 0.85.



(b) The top 15 features are first selected according to the F-value from χ^2 tests between labels (changer or no changer) and features. Then top 10 important features are listed after we permuted the features and fed them into the fitted random forest model. In *active morning* style, the ratio of morning to whole day is the most important feature.

Figure 11: Identifying changer v.s. no changer

4. Discussion and Conclusion

4.1. Targeting and Tailoring customers

In this research, we present a new approach for constructing dynamic energy lifestyles by applying LDA to residential electricity demand data.

Our framework is highly scalable and extensible, while also being flexible enough to accommodate different time intervals and completely new sources of residential energy data from other locations and contexts. Using this dynamic lifestyle approach, we can greatly simplify the interpretation of energy lifestyle patterns by using a method that generates a sparse number of energy attributes that are then used to generate a manageable set of energy lifestyle profiles. We show this process of generating energy lifestyles is robust to multiple load shape dictionary inputs and time intervals. We also demonstrate that these derived energy lifestyles can be associated with certain energy use characteristics, even though these energy use characteristics were not originally applied in constructing the lifestyles themselves.

We use these energy use characteristics to further interpret these lifestyles and provide insight into how such an approach toward lifestyle analysis could be used in practice. This energy lifestyle analysis approach can also be applied across different time horizons, allowing for applications at varying time intervals to examine temporal dynamics. While in our experiment we applied a seasonal time interval, shorter (e.g., monthly) or longer (e.g., yearly) time horizon energy lifestyles can also be estimated—dependent on data availability. Such an approach provides those who need to understand household energy lifestyles and lifestyle change patterns across time the ability to generate meaningful insights that can be applied to a wide variety of energy program designs and use cases. One example is exploring the demand flexibility for households. Such a demand flexibility is not only limited to considering electricity use timing throughout a day such as TOU programs [9, 10], but also reveals the day–to–day or even seasonal variations of energy use for households.

4.2. Potential applications of the dynamic lifestyles framework

We have identified three potential applications for this lifestyle analysis approach, each considering a different aspect of energy program design. The first application is in identifying households with lifestyle patterns that are most appropriate for installation of behind–the–meter resources, such as residential solar and battery storage systems. Taking the example of households with the energy lifestyle *home early*, these households may be particularly well–suited for rooftop residential solar as they have a pattern of usage that begins in midday, when solar energy potential is higher. A household where usage tends to peak later in the day and during evening hours with less solar energy potential, such as *home for dinner*, would be less suitable for targeting residential solar unless it was combined with battery storage (i.e., solar plus storage system) [52].

For demand response programs, certain energy lifestyles we derived from our experimental data suggest differing demand flexibility for households, especially when considering demand responsiveness to time-of-use pricing, which typically occurs during weekdays when system-level demand is highest, such as in the late afternoon and early evening. For households in *everyday is a new day*, their daily energy use is highly varied. This suggests that these households could be more able to change their daily energy use patterns, making them flexible in their demand because their energy use patterns are less structured compared to other energy lifestyles, such as *home early* and *home for dinner*. Energy lifestyles that are less flexible, such as *home early* and *home for dinner*, however, may be better suited for energy efficiency programs, because their lifestyle energy use patterns indicate stable electricity patterns with little day–to–day variation.

While both of these examples are related to households that display rel-

atively static energy lifestyle patterns, how these patterns differ across time is also important for potential applications in practice. First, if a household always displays a particular energy lifestyle pattern, this suggests that the household has a higher affinity toward the pattern of energy use within this lifestyle compared to a household that displays a change in lifestyle across time. Next, the number of lifestyle profile changes that a household undergoes on a seasonal basis, and the variety of these lifestyles, imparts important information about the household. Households that are constantly undergoing change will likely be difficult to target customers with demand response programs [33] given the instability of daily usage patterns. However, such a household may be a better candidate for an energy efficiency program, such as smart thermostat/AC.

While these examples have been targeted to energy provider applications, there are also opportunities to use this energy lifestyle analysis framework to inform energy intervention design, where households attempt to change their lifestyles to promote energy use patterns that save them money while also lessening their burden on the grid and carbon emissions. To do so, households that have a particular lifestyle with peak demand that corresponds with system demand, such as *home for dinner*, could attempt to change their usage to a different lifestyle pattern, such as *steady going* or *active morning*, with less usage concentrated during peak system demand periods. This energy lifestyle approach could then be used to determine if there is a shift in lifestyles, and also could become the basis in which to assess whether the household had successfully implemented this change. Moreover, such an approach may be used for households to quickly monitor their own energy lifestyle and make adjustments based on changes in the home or other new activity patterns [53]. In this respect, communicating information about en-

ergy use to customers via their lifestyle profile may be more impactful than other forms of more traditional energy use informational summaries (e.g., monthly kWh or energy cost).

Given the wide applications of this dynamic lifestyle approach, for which we have only provided a few examples, as well as the ability for iterative updating of energy lifestyles, we see great potential for building and extending this framework. However, our approach has some limitations. First, while we are able to verify these energy lifestyles using other energy use characteristics that were not included in the formulation of the energy lifestyles, we do not have an additional external measure to verify the presence or absence of this lifestyle based on other, non–energy–use information about household characteristics [43], such as number and age of occupants and patterns of activities in the home. Incorporating such data, if available, would be an important addition to this work and would bolster our framework's ability to provide insights about energy lifestyles. Additionally, the data that we applied in experiments to generate these energy lifestyles is from the early 2010s and therefore do not include recent trends in electricity use patterns within households related to smart home appliances, electric vehicles, and behind–the–meter resources [54], because these technologies were not yet widespread during this time period. To the extent the deployment of these technologies impact the formulation of these lifestyles themselves is not known, but we expect that solar, storage, and electric vehicles have some discernible impact on lifestyles that we do not capture here. At the same time, something like solar and storage could place a household in the *steady going* lifestyle, so while the formulation of the lifestyles may not be dramatically altered using more recent data, it could be that the proportion of households within a lifestyle will change to reflect new behind–the–meter

technology adoption.

4.3. Conclusion and next steps

We conceptualized and implemented a new approach for understanding energy lifestyles that can simplify interpretations about household energy use, has a high potential for applicability, can be easily scaled to larger datasets, and can measure changes in energy use across time. There are four immediate directions for future research as an extension to this work. First, this dynamic lifestyle approach can address a cold start problem in identifying patterns of use for new residential customers. Because this lifestyle approach can identify lifestyles using very sparse data inputs, energy providers could recommend energy program enrollment based on lifestyles after only the first few months of meter activation. Second, this dynamic lifestyle approach can be applied to additional residential datasets spanning different time periods and geographies to explore intra- and inter-yearly patterns in lifestyles as well as the influence of context and climate. Third, some steps of our lifestyles approach can incorporate privacy preserving methods, e.g. differential privacy [55, 56] or generative adversarial privacy [57, 58], to alleviate the concerns of revealing sensitive information of an individual household [59, 60], which is an important direction for future work. Lastly, using information about residential electricity data coupled with demographic and household characteristics, our model can further validate and provide new insights about lifestyles by identifying the characteristics related to different lifestyles and their dynamics across time.

Appendix A. Supplemental Material

Appendix A.1. Description of datasets

Our sampled homes covers eight different climate zones in California shown in Figure A.12.

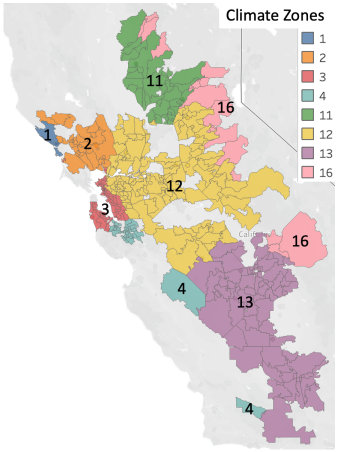


Figure A.12: Households are located in 8 climate zones in California, US.

Appendix A.2. Details of clustering load shapes

Appendix A.2.1. Clustering methods

We apply several clustering methods including k-means, k-medians, Hierarchical clustering, and DBSCAN for a thorough evaluation. For the center based methods (e.g. k-means and k-medians), we minimize the following objective (also known as distortion):

$$\min \sum_{i=1}^{N} d\big(\boldsymbol{x}_i, \phi(\boldsymbol{x}_i, C)\big), \tag{A.1}$$

where N is the sample size, d is the distance metric, and $\phi(\boldsymbol{x}_i, C)$ returns the nearest cluster center $c \in C$ to $\boldsymbol{x}_i$. When d is the Euclidean distance

and ϕ finds the nearest center using Euclidean distance, we have

$$\min \sum_{i=1}^{N} ||\boldsymbol{x}_i - \phi(\boldsymbol{x}_i, C)||_2^2 = \min \sum_{i=1}^{N} ||\boldsymbol{x}_i - \mu_{c^{(i)}}||_2^2 \quad , \tag{A.2}$$

where $c^{(i)}$ is the cluster label for i-th data point. To express the function ϕ more specifically, the k-means method updates the cluster centers by the following iterations until convergence:

$$c^{(i)} = \arg\min_j ||\boldsymbol{x}_i - \mu_j||_2, \quad \mu_j = \frac{\sum_{i=1}^{N} \mathbf{1}\{c^{(i)} = j\}\boldsymbol{x}_i}{\sum_{i=1}^{N} \mathbf{1}\{c^{(i)} = j\}} \quad . \tag{A.3}$$

The k-medians method differs from the previous k-means clustering when calculating the cluster center. Instead of taking the mean μ_j in equation (A.3), we compute the median as the center $\tilde{\mu}_j$ so that

$$\tilde{\mu}_j = median\{\boldsymbol{x}_{i=\{1\dots N\}}\}, \text{if } c^{(i)} = j, \forall i \in 1 \dots N \quad . \tag{A.4}$$

Hierarchical clustering is an agglomerative (hierarchical) approach, from the bottom individual point to up-level the whole dataset, that builds nested clusters in a successive manner [45, 61]. It has three popular implementations by minimizing different distances (objectives): Ward linkage [62], average linkage [63], and complete linkage [64]. The Ward's linkage method measures the distance between two clusters, A and B, which is how much the sum of squares will increase when we merge them:

$$\begin{aligned} \Delta(A, B) &= \sum_{i \in A \bigcup B} ||x_i - c_{A \cup B}||^2 - \sum_{i \in A} ||x_i - c_A||^2 - \sum_{i \in B} ||x_i - c_B||^2 \\ &= \frac{n_A n_B}{n_A + n_B} ||c_A - c_B||^2 \end{aligned} \tag{A.5}$$

where c_A, c_B are the centers of clusters A and B, and n_A, n_B are the number of points in clusters A and B. Δ denotes the merging cost of putting A and B together. The average linkage calculates the mean distance of all possible

pairs of points in two clusters. And the complete linkage method calculates the farthest distance of two points allocated in two clusters. In our setting, we pick Ward linkage because it gives a more stable result compared with other two types of linkage.

DBSCAN [46], known as density-based spatial clustering of applications with **n**oise, does not need to specify the number of clusters beforehand. It requires two key parameters, ϵ and $n_{\min}$, which define the neighborhood's distance and the minimum number of points to form a cluster. Higher $n_{\min}$ or lower ϵ indicate higher density to form a cluster. Choosing ϵ and $n_{\min}$ depends on domain knowledge of the data; hence, we evaluate multiple combinations and find it does not scale well for our use case.

Both Hierarchical and DBSCAN clustering do not compute cluster centers during iterations; therefore, we add an additional step to calculate a barycenter [65] of the points in each cluster to obtain a representative center. The barycenter is similar to the notion of a center in convex clusters, so we use the sequential averaging technique to compute the cluster center in the context of dealing with time series trajectories [66].

Appendix A.2.2. Evaluating different distances

To generate a robust and meaningful dictionary of load shapes, we compare several different distances such as cosine distance, L^1 distance (Manhattan distance), L^2 distance (Euclidean distance), and dynamic time warping (DTW). For simplicity of explanations, we consider two vectors $\boldsymbol{a}$ and $\boldsymbol{b} \in \mathbb{R}_+^m$ (e.g. $m = 24$) in the following context.

Cosine distance is a measure of similarity between two non-zero vectors

of an inner product space. The distance is expressed as

$$d_{cos}(\boldsymbol{a}, \boldsymbol{b}) = 1 - \frac{\boldsymbol{a}^T \boldsymbol{b}}{\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2} \quad . \tag{A.6}$$

L^1 distance is a measure of the element-wise absolute difference between two vectors. The expression is

$$d_{\ell_1}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_1 = \sum_{k=1}^{m} \big|\boldsymbol{a}[k] - \boldsymbol{b}[k]\big|, \tag{A.7}$$

where $\boldsymbol{a}[k]$ is the k-th dimension in vector $\boldsymbol{a}$.

L^2 distance is a measure of element-wise squared gap between two vectors. The expression is

$$d_{euc}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|_2 = \sqrt{\sum_{k=1}^{m} \big(\boldsymbol{a}[k] - \boldsymbol{b}[k]\big)^2}. \tag{A.8}$$

Dynamic Time Warping distance (DTW) is a method that calculates an optimal match between two given sequences [67]. We adopt a popular implementation that is based on dynamic programming:

$$d_{DTW}(\boldsymbol{a}, \boldsymbol{b}) = D(m, m), \text{ when } D(i, j) = \min \begin{cases} D(i-1, j) + \nu(i, j) \\ D(i-1, j-1) + \nu(i, j) \\ D(i, j-1) + \nu(i, j) \end{cases} , 1 \leq i, j \leq m \tag{A.9}$$

where D is a matrix that records the optimal warping value between the two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$; the $\nu(i, j)$ computes the cost between $\boldsymbol{a}[i]$ and $\boldsymbol{b}[j]$ (e.g. the cost is Euclidean distance in this example); and the base case is $D(0, 0) = (\boldsymbol{a}[0] - \boldsymbol{b}[0])^2$.

Hybrid distance: we additionally apply a mixture of the L^2 and DTW distances to compute the distance between two vectors:

$$d_{hybrid}(\boldsymbol{a}, \boldsymbol{b}) = \gamma d_{euc}(\boldsymbol{a}, \boldsymbol{b}) + (1 - \gamma) d_{DTW}(\boldsymbol{a}, \boldsymbol{b}), \tag{A.10}$$

where the $\gamma \in [0, 1]$ is the parameter to weigh the trade-off between two the distance metrics.

Appendix A.2.3. Evaluating clustering performances

To compare multiple clustering methods with different distances, we mainly use two evaluation metrics: *Calinski-Harabaz Index* [47] and *Davies-Bouldin Index* [68]. Both metrics are widely adopted to evaluate clustering models. A higher *Calinski-Harabasz Index* (CHI) relates to a model with better defined clusters, whereas a lower *Davies-Bouldin Index* (DBI) suggests to a model with a better separation between the clusters. To compare different clustering methods with various distances, we randomly draw 1000 data samples and record the cluster labels that yields the highest CHI and DBI scores when we search the number of clusters from $\{2, 4, 6, 8, 10, 12, 14, 16\}$. We repeat this exercise 5 times and present the results of the means of CHI and DBI in Table A.3.

Appendix A.2.4. Determining the dictionary size

Once the k-median method with the d_{hybrid} is chosen, we explore the appropriate size of the load shape dictionary. In particular, we tested the size of 100, 200, 300, 400, and 500 load shapes. Such a comparison involves two stages of clustering processes: 1) we randomly partition 60000 homes into 600 bins where each bin has 100 homes, and then we run clustering on these 100×365 data points for each bin to create 100 cluster centers. 2) Having these 100 clustered load shapes times the 600 bins, we run another round of clustering on 100×600 data points to yield the cluster centers with the size ranging from 100 to 500. Figure A.13 suggests that a size of 200 reduces the within-cluster distortion dramatically around 20%, which is

Table A.3: Clustering method comparison. We report the means of both *Calinksi-Harabaz Index* (CHI) and *Davies-Bouldin Index* (DBI) after 5 rounds of random tests. A higher CHI indicates a model can yield a better separation of clusters. In contrast, a lower DBI suggests a better separation between the clusters. We see that k-medians with the hybrid distance gives the best clustering performance.

method	distance	$CHI \uparrow$	$DBI \downarrow$
k-means	Euclidean	107.42	4.53
	cosine	102.31	3.87
	ℓ_1	99.51	4.14
	DTW	113.93	3.89
	$d_{hybrid}(\gamma = 0.5)$	116.76	3.67
k-median	Euclidean	109.53	4.50
	cosine	108.11	4.05
	ℓ_1	102.40	4.19
	DTW	115.84	3.82
	$\boldsymbol{d_{hybrid}(\gamma = 0.5)}$	**118.31**	**3.54**
Hierarchical (Ward)	Euclidean	93.21	4.99
	cosine	92.18	4.81
	ℓ_1	90.53	5.16
	DTW	98.65	4.87
	$d_{hybrid}(\gamma = 0.5)$	101.32	4.58
DBSCAN ($\epsilon = 0.1$)	Euclidean	82.44	5.17
	cosine	85.37	5.29
	ℓ_1	80.15	5.18
	DTW	88.03	5.25
	$d_{hybrid}(\gamma = 0.5)$	89.75	5.07

much more prominent than at other sizes. Thus, we pick 200 clusters as the size of the load shape dictionary.

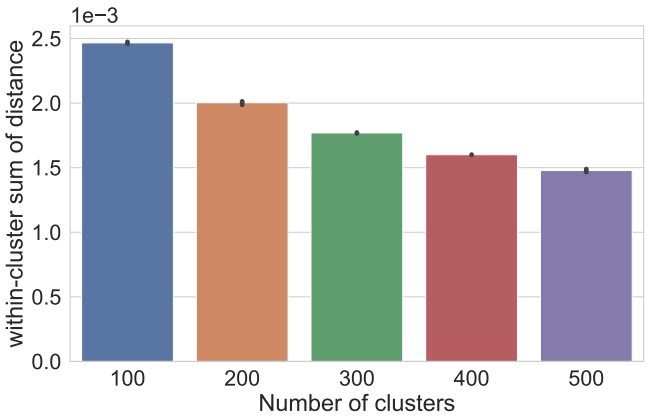


Figure A.13: Choosing the size of load shape dictionary. When increasing the number of clusters above 200, we have limited marginal gain of reducing the within cluster sum of distances. Thus, we consider 200 as the proper dictionary size.

Appendix A.3. Generating distinct attributes

Before synthesizing the energy lifestyles of households, we need to find the representative attributes that compose the multiple load patterns for households. Thus, teasing out distinct latent attributes of energy usage is a crucial building block. We use the LDA with a prescribed $K = 10$ number of attributes (topics), displayed in Figure A.14. After fitting the LDA, we find several attributes are very similar to each other such as *attribute 1* and *attribute 6* in Figure A.14. A further calculation of the correlation distances between attributes (normalized 24-dimensional vectors) also demonstrates that some attributes are very close and can be merged together (Figure A.15), where the correlation distances between two vectors $\boldsymbol{a}$

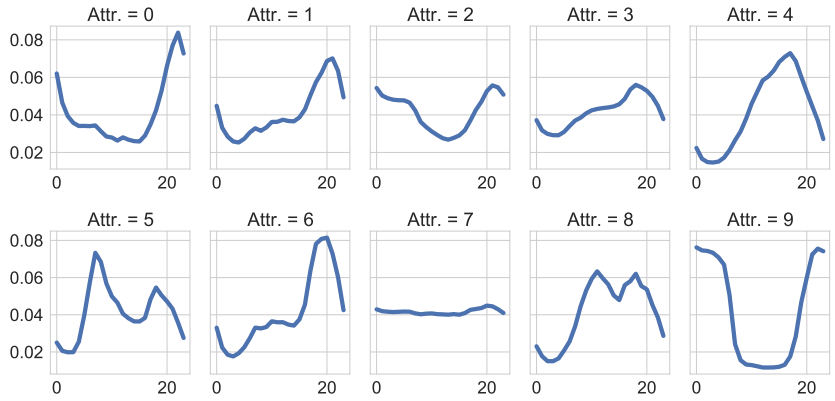


Figure A.14: Ten attributes are obtained after applying LDA initially. A few center curves are similar, such as Attribute 1 and Attribute 6. We then construct a projection matrix according to correlation distance to reduce the number of attributes (a.k.a, number of topics) down to six.

and $\boldsymbol{b}$ with their associated elements means $\mu_{\boldsymbol{a}}$ and $\mu_{\boldsymbol{b}}$ can be expressed as

$$d_{corr} = 1 - \frac{(\boldsymbol{a} - \mu_{\boldsymbol{a}})(\boldsymbol{b} - \mu_{\boldsymbol{b}})}{\|\boldsymbol{a} - \mu_{\boldsymbol{a}}\|_2 \|\boldsymbol{b} - \mu_{\boldsymbol{b}}\|_2}. \tag{A.11}$$

We set the threshold as 0.1 to indicate that two attributes are very similar, and then find the nearest neighbors of the energy attributes based on that criterion. Once the neighbors are settled, we merge similar shapes together by 1) constructing a projection matrix $A^T A$ where $A = D_{corr} + I$ and where D_{corr} consists of either zeros or ones, where ones mean when d_{corr} is less than 0.1 in entries, mentioned in equation (A.11) and I is the identity matrix; 2) scanning through columns and pruning the $A^T A$ once the corresponding rows are located. In our experiment, we prune down to 6 dimensions, because each dimension has its distinct attribute shape (Figure 2). Additionally, we qualitatively verify that 6 attributes are robust for a large population by randomly sampling 2000 homes and comparing their corre-

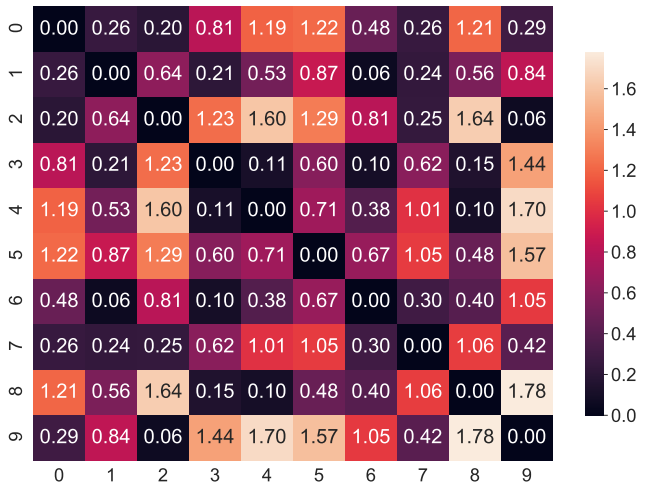


Figure A.15: Correlation distance heatmap of ten attributes that are obtained from Figure A.14

lation distances on the attribute spaces prior of projection (Figure A.16). We observe that homes are nested mainly into 5 to 6 diagonal blocks, which supports our previous merge operation of simplifying the energy attributes.

Having determined the energy attributes, we use the 6-dimensional vector to represent each home. In order to obtain prototypical attributes distribution of these homes, we need to segment all the homes using another round of clustering. We use k-means with $K = 6$, because this setting gives a distinguishable and meaningful result. The corresponding centers of the attribute weights are shown in Figure A.17.

Appendix A.4. Population change over seasons

We provide detailed population splits of six lifestyles over four seasons in Table A.4. The numbers are the counts, and percentage values in the

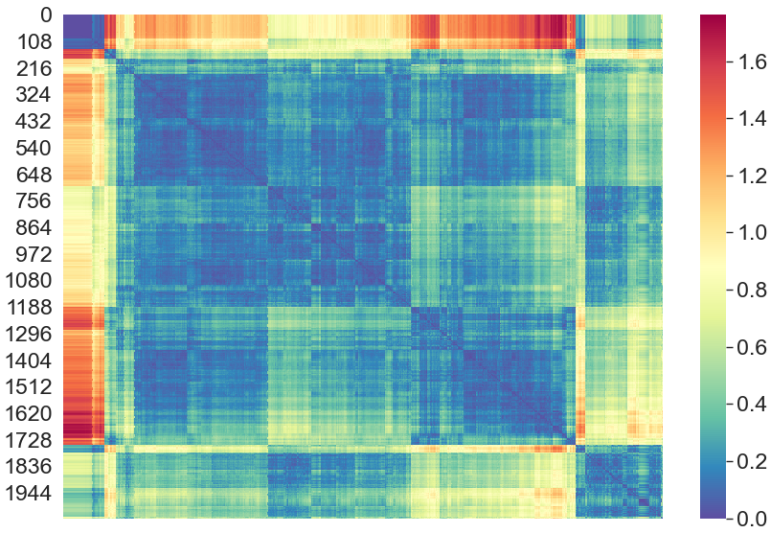


Figure A.16: Distance heatmap. 2000 homes are randomly sampled and their pairwise correlation distances appear to be segmented into 6 main blocks along the diagonal.

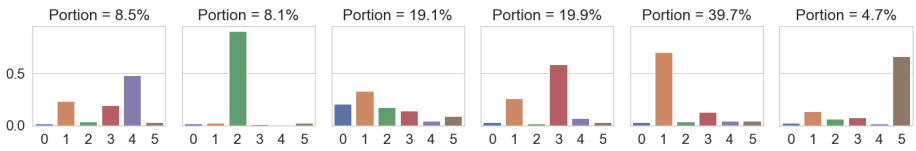


Figure A.17: Weights of energy attributes

parentheses are proportions of the population. From the table, we find that the lifestyle of home for dinner is the most frequently occurring, usually accounting for about 40% of the households except in the autumn when it accounts for 33.02% of the samples.

Table A.4: Population split of lifestyles in seasons

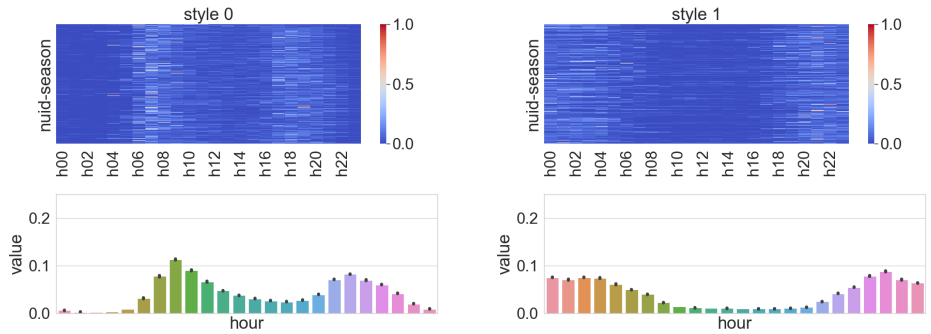
	Autumn	**Winter**	**Spring**	**Summer**
active morning	4027 (6.71%)	7836 (13.06%)	4713 (7.85%)	3968 (6.61%)
night owl	5174 (8.62%)	4844 (8.07%)	5504 (9.17%)	4557 (7.60%)
everyday is a new day	8632 (14.39%)	5509 (9.18%)	10750 (17.92%)	21311 (35.52%)
home early	18963 (31.61%)	11975 (19.96%)	10254 (17.09%)	4420 (7.37%)
home for dinner	19813 (33.02%)	26084 (43.47%)	24458 (40.76%)	25744 (42.91%)
steady going	3391 (5.65%)	3752 (6.25%)	4221 (7.04%)	0 (0%)[†]

[†] We do not observe that households in our samples have a flat pattern of energy use (i.e., steady going lifestyle) across many days in the summer.

Appendix A.5. Features of energy usage

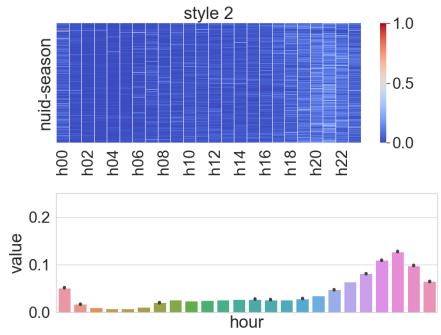
We show distributions of additional features associated with different lifestyles. The definitions of features are provided in Table 2.

First, we provide the peak hour distribution for the *home early* lifestyle in addition to the other styles mentioned in Figure A.18. Second, multiple year-specific features are displayed in Figure A.20 over six lifestyles. Finally, we also include additional plots describing the seasonal features. Figure A.21–A.24 demonstrate the stability of the group of No Changers. These figures cover the different distributions of No Changers across four seasons including morning energy use, evening energy use, peak energy use, and hourly average energy use.

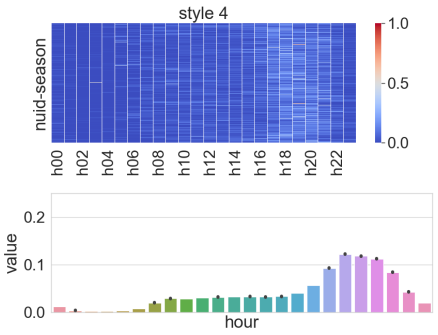


(a) active morning

(b) night owl



(c) everyday is a new day



(d) home for dinner

Figure A.18: Peak hour distribution over a day (hour 0 – hour 23). In each sub-figure, the upper panel shows the heatmap of peak hour frequency when each home in a season is represented by each row stacked by seasons. The lower panel is the averaged frequency of peak hour occurrence for all homes in the corresponding lifestyle group.

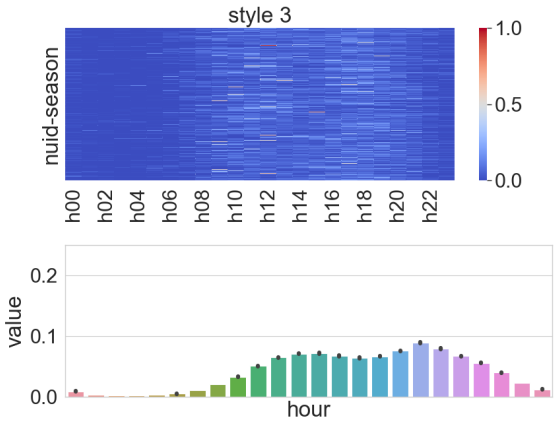


Figure A.19: Peak hour distributions of *home early* lifestyle.

To provide the detailed comparisons between Changers and No Changers, we show the ratio of night to whole day usage and the ratio of noon to whole day usage in Figure A.25 and Figure A.26. Because the distributions of those two features significantly reveal the seasonal variations for the Changer group.

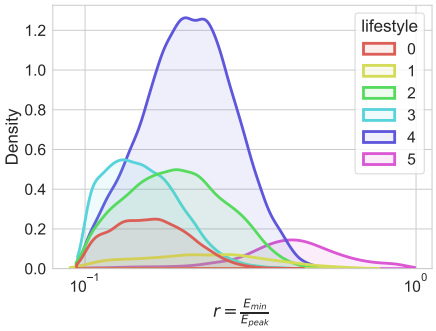
Appendix A.6. Classification details

Appendix A.6.1. Identifying lifestyles

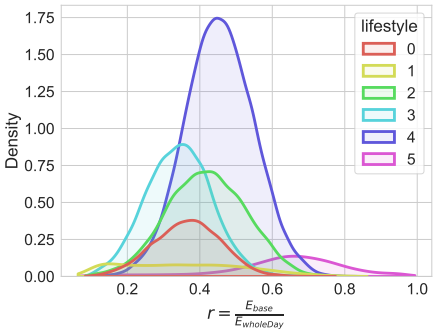
We provide the performance details of classifying lifestyles using random forest fed with load features. The *night owl* has the highest F1 score around 0.84. In contrast, the *home early* lifestyle has the lowest F1 score about 0.55, indicating this is a difficult lifestyle to identify.

Appendix A.6.2. Identifying no changer

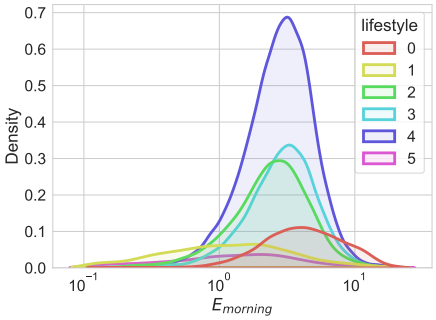
We classify Changer vs No Changer in each lifestyle. Because the summer season does not have a *steady going* group, we show the other five lifestyles when each of them has a group of No Changers over four seasons



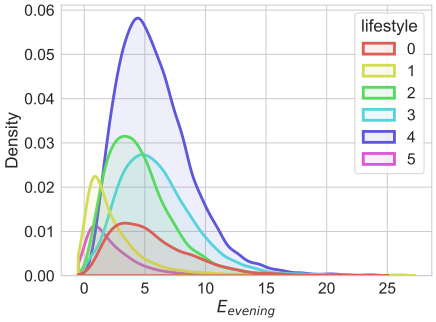
(a) Ratio of hourly min- over max-energy in 24 hours



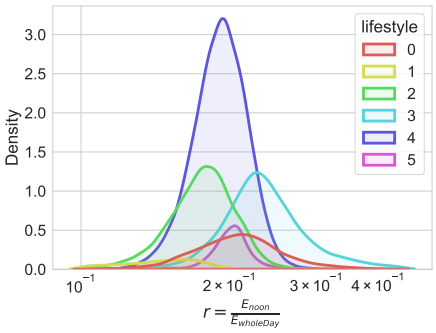
(b) Ratio of base over whole day energy use in 24 hours



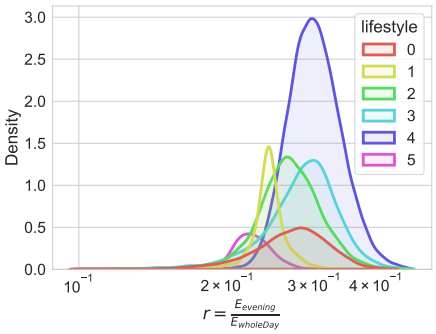
(c) Morning energy use (6am - 10am)



(d) Evening energy use (6pm-10pm)



(e) Ratio of noon over whole day energy use



(f) Ratio of evening over whole day energy use

Figure A.20: Distributions of different energy usage features characterizing the distinctions between lifestyles.

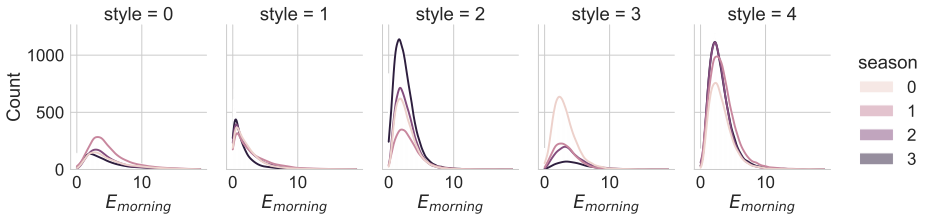


Figure A.21: Distribution of morning energy use (in KWh) over four seasons for lifestyles

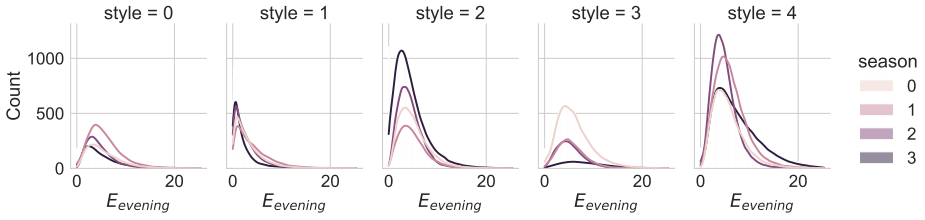


Figure A.22: Distribution of evening energy use (in KWh) over four seasons for lifestyles

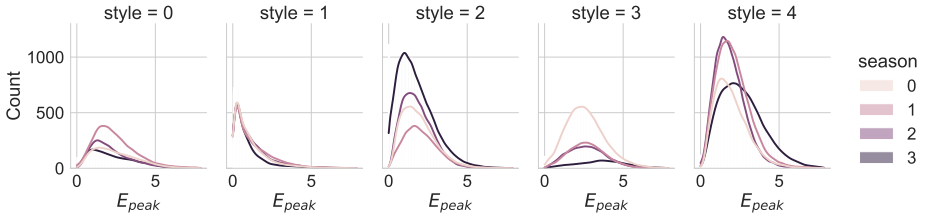


Figure A.23: Distribution of daily peak energy (in KWh) over four seasons for lifestyles

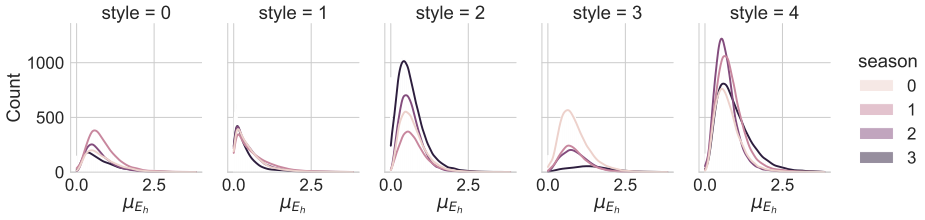


Figure A.24: Distribution of hourly mean energy (in KWh) over four seasons for lifestyles

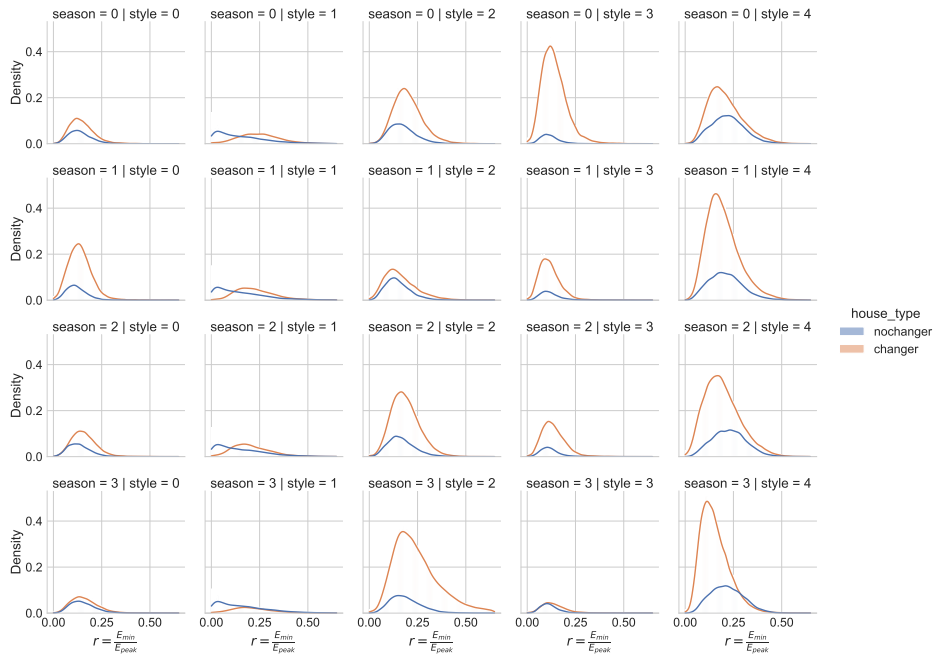


Figure A.25: Distributions of min (base) to peak energy ratio for Changers and No Changers of five lifestyles over four seasons.

Table A.5: Lifestyle classification performance

style index	lifestyle	precision	recall	F1 score
0	active morning	0.7164	0.6315	0.6713
1	night owl	0.8657	0.8176	0.8409
2	everyday is a new day	0.6411	0.6191	0.6299
3	home early	0.6551	0.4685	0.5463
4	home for dinner	0.6652	0.7841	0.7198
5	steady going	0.6417	0.5493	0.5919
				average acc = 0.685

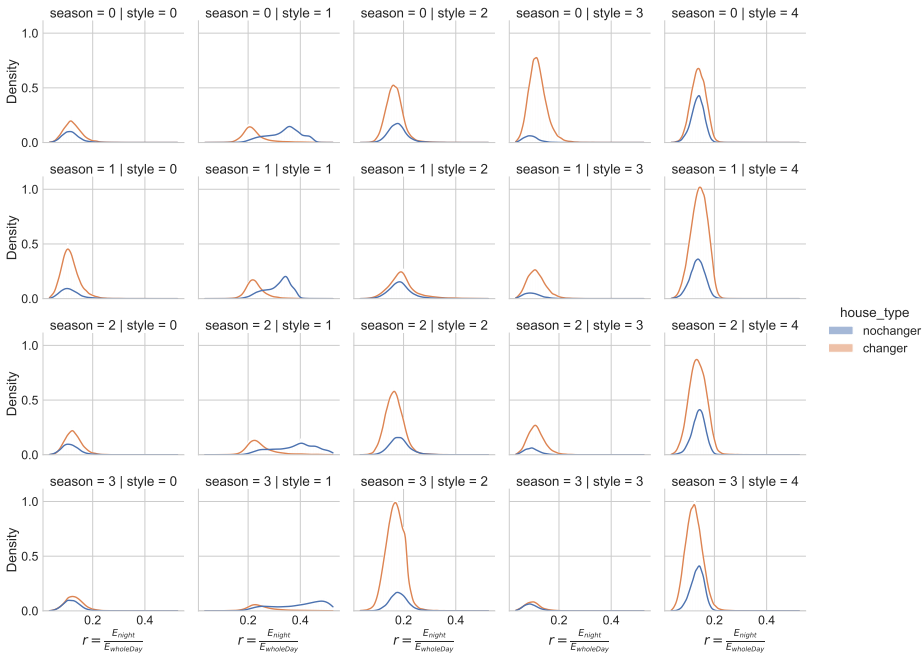


Figure A.26: Distributions of night to whole day energy ratio for Changers and No Changers of five lifestyles over four seasons.

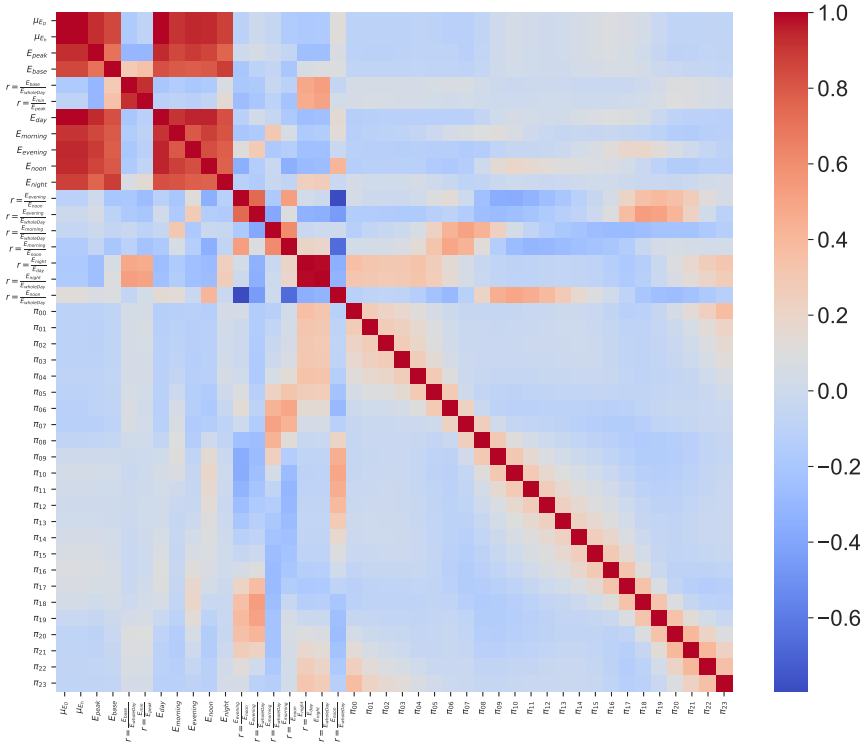


Figure A.27: Correlation heatmap of features.

(Table A.6, Table A.7, Table A.8, Table A.9, and Table A.10). We observe that identifying a Changer is generally easier than identifying a No Changer because of the higher F1 scores. One exception is the *night owl* lifestyle, which has similar performances of identifying Changers and No Changers given their relatively similar F1 scores. In addition, we show AUC plots identifying Changers vs No Changers for those five lifestyles (Figure A.28).

The most important determinants of identifying Changers or No changers are different across the 5 lifestyles. We present their corresponding top 10 important features in Figure A.29.

Table A.6: active morning lifestyle

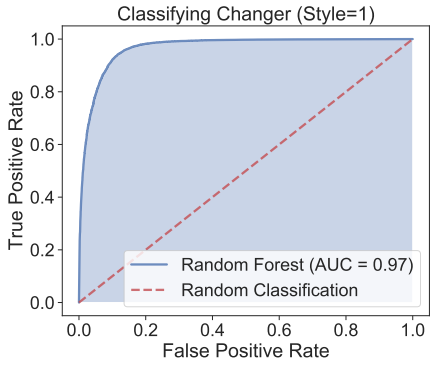
label	precision	recall	F1 score
No Changer (0)	0.6481	0.2502	0.3611
Changer (1)	0.8922	0.9786	0.9334
			average acc = 0.879

Table A.7: night owl lifestyle

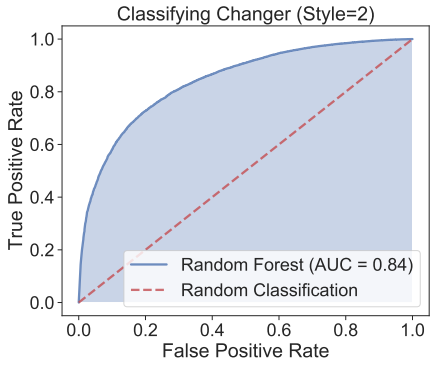
label	precision	recall	F1 score
No Changer (0)	0.9301	0.8706	0.8994
Changer (1)	0.9073	0.9508	0.9285
			average acc = 0.917

Appendix A.7. Exploration on influencing features

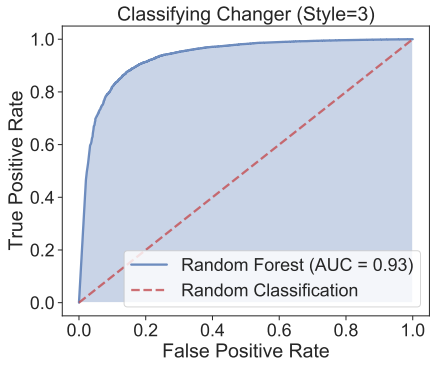
To verify our identified lifestyles alternatively, we run multinomial logit model to measure how the electricity features influence the (log) odds of



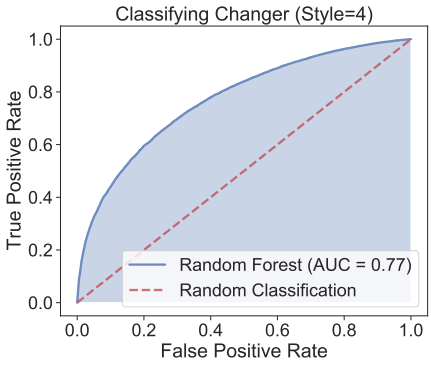
(a) *night owl* style



(b) *everyday is a new day* style



(c) *home early* style



(d) *home for dinner* style

Figure A.28: AUC of classifying Changer v.s. No Changer.

Table A.8: everyday is a new day lifestyle

label	precision	recall	F1 score
No Changer (0)	0.6476	0.1533	0.2479
Changer (1)	0.9083	0.9902	0.9475
			average acc = 0.902

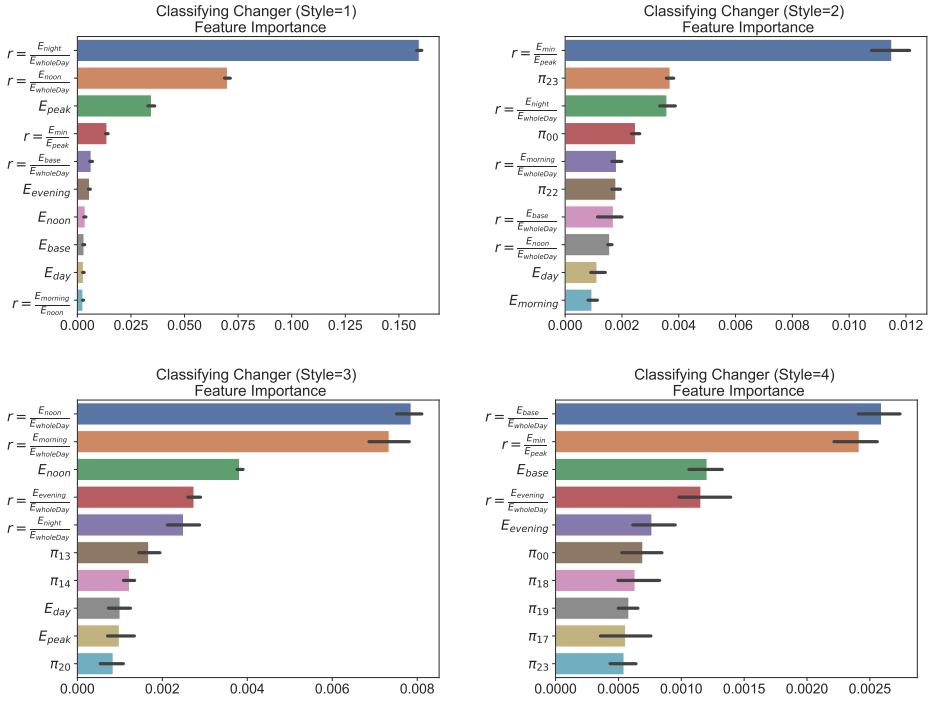


Figure A.29: Feature importance (Changer v.s. No changer)

Table A.9: home early lifestyle

label	precision	recall	F1 score
No Changer (0)	0.6966	0.4028	0.5104
Changer (1)	0.9657	0.9897	0.9775
			average acc = 0.957

Table A.10: home for dinner lifestyle

label	precision	recall	F1 score
No Changer (0)	0.5919	0.1734	0.2682
Changer (1)	0.9657	0.9897	0.9775
			average acc = 0.856

having a certain lifestyle among our experimental samples. We treat the steady going lifestyle as the reference group, without loss of the generality, setting it the reference group $Y = 0$. Thus the model is comparing each group outcome with the reference group:

$$\log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \boldsymbol{\theta}_1\boldsymbol{x}, \cdots, \log\left(\frac{P(Y=5)}{P(Y=0)}\right) = \boldsymbol{\theta}_5\boldsymbol{x}. \tag{A.12}$$

where $Y = 1, ..., 5$ indicates Active morning, Night owl, Everyday is a new day, Home early, and Home for dinner respectively. The linear coefficients $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_5$ are associated with the lifestyles and the input features are denoted as $\boldsymbol{x}$. Notice that we normalize input features between 0 and 1 so that all the scales are in the similar range. We make use of the statsmodels [69] [2]

[2] https://www.statsmodels.org/stable/index.html

package and use L^1 regularization to fit the parameters and output p-values and 95% generate confidence intervals. The results are averaged values from 150 iterations with batch size of 5000.

We mainly focus on the features of energy consumption volumes and consumption ratios within a day, because they are direct proxy features of daily activities. In the following comparison, we consider base energy, base to peak ratio, base energy portion (relative to a day), ratio of evening to whole day, ratio of morning to whole day, ratio of night to whole day, and ratio of noon to whole day, since they are straightforward to harness with daily routines. Table A.11 to Table A.15 present the results.

Table A.11: Multinomial logistic regression results: Active morning.

Param. †	**coef**	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
base_ene	39.9405	16.7984	2.3530	0.122	7.016	72.865
bp_ratio	-12.6192	2.8126	-4.4661	0.001	-18.132	-7.107
base_portion	-0.4703	2.1370	-0.2226	0.449	-4.659	3.718
r_e2w	40.0273	3.7593	10.6363	0.000	32.659	47.395
r_m2w	45.5967	3.8572	11.8167	0.000	38.037	53.157
r_ni2w	-23.5468	4.4598	-5.2854	0.000	-32.288	-14.806
r_no2w	19.0893	3.9245	4.8517	0.001	11.397	26.781
const	-15.8458	2.0431	-7.7429	0.000	-19.850	-11.842

† The parameters from top to bottom rows are base energy, base to peak ratio, base energy portion (relative to a day), ratio of evening to whole day, ratio of morning to whole day, ratio of night to whole day, ratio of noon to whole day, and the constant term.

Table A.12: Multinomial logistic regression results: Night owl.

Param.	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
base_ene	29.3953	13.1345	2.1403	0.126	3.652	55.138
bp_ratio	-2.8727	1.9517	-1.4376	0.255	-6.698	0.953
base_portion	-2.3137	1.9372	-1.1984	0.325	-6.111	1.483
r_e2w	33.0589	4.6545	7.0663	0.000	23.936	42.182
r_m2w	13.5410	4.0855	3.2894	0.031	5.533	21.548
r_ni2w	39.3345	4.7984	8.1428	0.000	29.930	48.739
r_no2w	-12.9092	5.0904	-2.6113	0.130	-22.886	-2.932
const	-13.6910	2.8082	-4.7628	0.001	-19.195	-8.187

Table A.13: Multinomial logistic regression results: Everyday is a new day.

Param.	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
base_ene	-6.2529	12.7474	-0.6382	0.344	-31.237	18.731
bp_ratio	-12.2651	1.8272	-6.6772	0.000	-15.846	-8.684
base_portion	3.2624	1.7237	1.8944	0.165	-0.116	6.641
r_e2w	27.9164	2.5287	11.0360	0.000	22.960	32.873
r_m2w	-10.0360	3.2348	-3.1260	0.029	-16.376	-3.696
r_ni2w	5.4709	2.8215	1.9058	0.131	-0.059	11.001
r_no2w	-3.0341	2.8927	-1.0808	0.346	-8.704	2.636
const	-1.4733	0.5860	-2.4375	0.054	-2.622	-0.325

Table A.14: Multinomial logistic regression results: Home early.

Param.	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
base_ene	63.0495	15.3862	4.0614	0.003	32.893	93.206
bp_ratio	-25.1830	2.7710	-9.0725	0.000	-30.614	-19.752
base_portion	5.0066	2.0665	2.4281	0.115	0.956	9.057
r_e2w	32.1368	3.0962	10.3711	0.000	26.068	38.205
r_m2w	-1.7445	3.4999	-0.5145	0.448	-8.604	5.115
r_ni2w	-24.0826	3.7348	-6.4682	0.000	-31.403	-16.762
r_no2w	30.1940	3.4867	8.6466	0.000	23.360	37.028
const	-6.5285	1.3256	-4.8937	0.000	-9.127	-3.930

Table A.15: Multinomial logistic regression results: Home for dinner.

Param.	**coef**	**std err**	**z**	**P> |z|**	**[0.025**	**0.975]**
base_ene	41.2951	12.0066	3.3187	0.016	17.763	64.828
bp_ratio	-5.9587	1.8252	-3.2359	0.028	-9.536	-2.381
base_portion	1.1285	1.7586	0.6460	0.387	-2.318	4.575
r_e2w	50.2073	3.0536	16.4404	0.000	44.222	56.192
r_m2w	2.2251	3.3439	0.6512	0.413	-4.329	8.779
r_ni2w	-13.5150	3.4060	-3.9894	0.006	-20.191	-6.839
r_no2w	19.6216	3.3668	5.8139	0.000	13.023	26.220
const	-11.8697	1.3205	-8.9764	0.000	-14.458	-9.282

Appendix A.8. Description of Latent Dirichlet Allocation

In this section, we describe details of Latent Dirichlet Allocation (LDA) and its application in constructing lifestyles. We use the notation listed in Table A.16 for the remaining part.

Notation	Description
k	Index of attributes (topics)
K	Number of attributes
i	Index of shapes
j	Index of homes or users
α	Dirichlet prior on the attributes in a home
β	Dirichlet prior weight of shapes in a attribute
θ_j	Attribute distribution of home j
θ_{jk}	Proportion of attribute k in home j
ψ_k	Shape distribution of attribute k
ψ_{ki}	Probability of word i occurring in attribute k
s_j	Shape collection of home j
s_{ji}	shape i in s_j
z_{ji}	Attribute assignment for shape s_{ji} from home j
M	Number of homes
N_j	Number of shapes in home j

Table A.16: LDA symbol description

The LDA model first prescribes K attributes, with each attribute k associated with a distribution ψ_k over shapes in the dictionary. In particular, ψ_k is sampled from a Dirichlet distribution $Dir(\beta)$. Based on these created attributes, a home j (namely a collection of shapes s_j) is generated by first

sampling a distribution θ_j over K attributes from another Dirichlet distribution $Dir(\alpha)$, which determines attribute assignment for each shape in s_j and then choosing each shape s_{ji} based on θ_j. In generating each shape s_{ji}, LDA first samples a particular attribute $z_{ji} \in \{1, ..., K\}$ from multinomial distribution θ_j, and then the shape s_{ji} is selected from a multinomial distribution $\psi_{z_{ji}}$. This process can be summarized to the following steps:

Steps in Latent Dirichlet Allocation

step1: pick shape distribution of each attribute k by $\psi_k \sim Dir(\beta)$

step2: pick attribute distribution for each home j by $\theta_j \sim Dir(\alpha)$

step3: For each home j, for each shape s_{ji} in j:

pick an attribute $z_{ji} \sim \theta_j$;

pick a shape $s_{ji} \sim \psi_{z_{ji}}$

The model inference can be done by using variational expectation-maximization (EM) [41, 36] or Markov Chain Monte Carlo methods (e.g. Gibbs sampling [70]). Both methods can infer the posterior of attribute distribution θ and attribute-shape distribution ψ efficiently. In our experiment, we use sklearn [71] with variational EM algorithm[3] to perform the computation.

References

[1] A. Ghosal, M. Conti, Key management systems for smart grid advanced metering infrastructure: A survey, IEEE Communications Surveys Tutorials 21 (3) (2019) 2831–2848. doi:10.1109/COMST.2019.2907650.

[3] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[2] F. Fahiman, S. M. Erfani, S. Rajasegarar, M. Palaniswami, C. Leckie, Improving load forecasting based on deep learning and k-shape clustering, in: 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 4134–4141. doi:10.1109/IJCNN.2017.7966378.

[3] E. Barbour, M. González, Enhancing household-level load forecasts using daily load profile clustering, in: Proceedings of the 5th Conference on Systems for Built Environments, BuildSys '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 107–115. doi:10.1145/3276774.3276793.
URL https://doi.org/10.1145/3276774.3276793

[4] B. Yildiz, J. I. Bilbao, J. Dore, A. Sproul, Household electricity load forecasting using historical smart meter data with clustering and classification techniques, in: 2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia), 2018, pp. 873–879. doi:10.1109/ISGT-Asia.2018.8467837.

[5] J. Kwac, R. Rajagopal, Data-driven targeting of customers for demand response, IEEE Transactions on Smart Grid 7 (5) (2016) 2199–2207. doi:10.1109/TSG.2015.2480841.

[6] J. Wong, R. Rajagopal, A simple way to use interval data to segment residential customers for energy efficiency and demand response program targeting, in: ACEEE Proceedings, 2012, pp. 1–9.

[7] A. Todd-Blick, C. A. Spurlock, L. Jin, P. Cappers, S. Borgeson, D. Fredman, J. Zuboy, Winners are not keepers: Characterizing household engagement, gains, and energy patterns in demand response using ma-

chine learning in the united states, Energy Research & Social Science 70 (2020) 101595.

[8] A. Kavousian, R. Rajagopal, M. Fischer, Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior, Energy 55 (2013) 184–194.

[9] V. Di Cosmo, D. O'Hora, Nudging electricity consumption using tou pricing and feedback: evidence from irish households, Journal of Economic Psychology 61 (2017) 1–14.

[10] M. Afzalan, F. Jazizadeh, Residential loads flexibility potential for demand response using energy consumption patterns and user segments, Applied Energy 254 (2019) 113693.

[11] I. Dusparic, A. Taylor, A. Marinescu, F. Golpayegani, S. Clarke, Residential demand response: Experimental evaluation and comparison of self-organizing techniques, Renewable and Sustainable Energy Reviews 80 (2017) 1528–1536.

[12] H. Boudet, C. Zanocco, G. Stelmach, M. Muttaqee, J. Flora, Public preferences for five electricity grid decarbonization policies in california, Review of Policy Research (2021).

[13] L. Werner, A. Wierman, S. H. Low, Pricing flexibility of shiftable demand in electricity markets, in: Proceedings of the Twelfth ACM International Conference on Future Energy Systems, 2021, pp. 1–14.

[14] J. Kwac, R. Rajagopal, Demand response targeting using big data an-

alytics, in: 2013 IEEE International Conference on Big Data, 2013, pp. 683–690. doi:10.1109/BigData.2013.6691643.

[15] J. Kwac, J. Flora, R. Rajagopal, Lifestyle segmentation based on energy consumption data, IEEE Transactions on Smart Grid 9 (4) (2018) 2409–2418. doi:10.1109/TSG.2016.2611600.

[16] O. Motlagh, A. Berry, L. O'Neil, Clustering of residential electricity customers using load time series, Applied energy 237 (2019) 11–24.

[17] I. Abubakar, S. Khalid, M. Mustafa, H. Shareef, M. Mustapha, Application of load monitoring in appliances' energy management–a review, Renewable and Sustainable Energy Reviews 67 (2017) 235–245.

[18] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.

[19] M. Hoffman, F. Bach, D. Blei, Online learning for latent dirichlet allocation, advances in neural information processing systems 23 (2010) 856–864.

[20] P. Pinoli, D. Chicco, M. Masseroli, Latent dirichlet allocation based on gibbs sampling for gene function prediction, in: 2014 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, 2014, pp. 1–8. doi:10.1109/CIBCB.2014.6845514.

[21] M. Lienou, H. Maitre, M. Datcu, Semantic annotation of satellite images using latent dirichlet allocation, IEEE Geoscience and Remote Sensing Letters 7 (1) (2010) 28–32. doi:10.1109/LGRS.2009.2023536.

[22] P. Grunewald, M. Diakonova, The electricity footprint of household

activities-implications for demand models, Energy and Buildings 174 (2018) 635–641.

[23] E. Aydin, D. Brounen, N. Kok, Information provision and energy consumption: Evidence from a field experiment, Energy Economics 71 (2018) 403–410.

[24] P. M. Gladhart, B. M. Morrison, J. J. Zuiches, Energy and families: Lifestyles and energy consumption in Lansing, Michigan State University Press, 1987.

[25] L. Lutzenhiser, Behavioral assumptions underlying california residential sector energy efficiency programs, UC Berkeley: California Institute for Energy and Environment (2009).

[26] J. Torriti, Understanding the timing of energy demand through time use data: Time of the day dependence of social practices, Energy research & social science 25 (2017) 37–47.

[27] B. Anderson, Laundry, energy and time: Insights from 20 years of time-use diary data in the united kingdom, Energy Research & Social Science 22 (2016) 125–136.

[28] A. Warde, S.-L. Cheng, W. Olsen, D. Southerton, Changes in the practice of eating: A comparative analysis of time-use, Acta Sociologica 50 (4) (2007) 363–385. doi:10.1177/0001699307083978.

[29] D. Southerton, Analysing the temporal organization of daily life: Social constraints, practices and their allocation, Sociology 40 (3) (2006) 435–454.

[30] T. Memmott, S. Carley, M. Graff, D. M. Konisky, Sociodemographic disparities in energy insecurity among low-income households before and during the covid-19 pandemic, Nature Energy 6 (2) (2021) 186–193.

[31] L. M. Giurge, A. V. Whillans, A. Yemiscigil, A multicountry perspective on gender differences in time use during covid-19, Proceedings of the National Academy of Sciences 118 (12) (2021).

[32] C. Zanocco, J. Flora, R. Rajagopal, H. Boudet, Exploring the effects of california's covid-19 shelter-in-place order on household energy practices and intention to adopt smart home technologies, Renewable and Sustainable Energy Reviews (2020) 110578.

[33] A. Todd, P. Cappers, C. A. Spurlock, L. Jin, Spillover as a cause of bias in baseline evaluation methods for demand response programs, Applied Energy 250 (2019) 344–357.

[34] R. H. Socolow, The twin rivers program on energy conservation in housing: Highlights and conclusions, Energy and Buildings 1 (3) (1978) 207–242.

[35] M. Gerlach, T. P. Peixoto, E. G. Altmann, A network approach to topic models, Science advances 4 (7) (2018) eaaq1360.

[36] D. M. Blei, Probabilistic topic models, Communications of the ACM 55 (4) (2012) 77–84.

[37] J. J. J. van Der Hooft, J. Wandy, M. P. Barrett, K. E. Burgess, S. Rogers, Topic modeling for untargeted substructure exploration

in metabolomics, Proceedings of the National Academy of Sciences 113 (48) (2016) 13738–13743.

[38] M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference., Journal of Machine Learning Research 14 (5) (2013).

[39] T. Salimans, D. Kingma, M. Welling, Markov chain monte carlo and variational inference: Bridging the gap, in: International Conference on Machine Learning, PMLR, 2015, pp. 1218–1226.

[40] L. Giordono, H. Boudet, A. Gard-Murray, Local adaptation policy responses to extreme weather events, Policy sciences 53 (4) (2020) 609–636.

[41] D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, 2006, pp. 113–120.

[42] R. K. Jain, J. Qin, R. Rajagopal, Data-driven planning of distributed energy resources amidst socio-technical complexities, Nature Energy 2 (8) (2017) 1–11.

[43] C. Beckel, L. Sadamori, T. Staake, S. Santini, Revealing household characteristics from smart meter data, Energy 78 (2014) 397–410.

[44] P. S. Bradley, O. L. Mangasarian, W. N. Street, Clustering via concave minimization, Advances in neural information processing systems (1997) 368–374.

[45] S. C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (3) (1967) 241–254.

[46] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226–231.

[47] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 3 (1) (1974) 1–27.

[48] S. Xu, E. Barbour, M. C. González, Household segmentation by load shape and daily consumption, in: Proc. of. ACM SigKDD Workshop, 2017, pp. 1–9.

[49] L. Breiman, Random forests, Machine learning 45 (1) (2001) 5–32.

[50] A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously., Journal of Machine Learning Research 20 (177) (2019) 1–81.

[51] J. M. Abreu, F. C. Pereira, P. Ferrão, Using pattern recognition to identify habitual behavior in residential electricity consumption, Energy and buildings 49 (2012) 479–487.

[52] A. Pena-Bello, E. Barbour, M. Gonzalez, M. Patel, D. Parra, Optimized pv-coupled battery systems for combining applications: Impact of battery technology and geography, Renewable and Sustainable Energy Reviews 112 (2019) 978–990.

[53] A. Albert, R. Rajagopal, Smart meter driven segmentation: What your consumption says about you, IEEE Transactions on power systems 28 (4) (2013) 4019–4030.

[54] J. P. Carvallo, S. Bieler, M. Collins, J. Mueller, C. Gehbauer, D. J. Gotham, P. H. Larsen, A framework to measure the technical, economic, and rate impacts of distributed solar, electric vehicles, and storage, Applied Energy 297 (2021) 117160.

[55] C. Dwork, Differential privacy: A survey of results, in: International conference on theory and applications of models of computation, Springer, 2008, pp. 1–19.

[56] C. Dwork, A. Roth, et al., The algorithmic foundations of differential privacy., Found. Trends Theor. Comput. Sci. 9 (3-4) (2014) 211–407.

[57] C. Huang, P. Kairouz, X. Chen, L. Sankar, R. Rajagopal, Context-aware generative adversarial privacy, Entropy 19 (12) (2017) 656.

[58] X. Chen, T. Navidi, R. Rajagopal, Generating private data with user customization, arXiv preprint arXiv:2012.01467 (2020).

[59] X. Chen, P. Kairouz, R. Rajagopal, Understanding compressive adversarial privacy, in: 2018 IEEE Conference on Decision and Control (CDC), IEEE, 2018, pp. 6824–6831.

[60] X. Chen, T. Navidi, R. Rajagopal, Energy resource control via privacy preserving data, Electric Power Systems Research 189 (2020) 106719.

[61] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of Massive Datasets, 2nd Edition, Cambridge University Press, USA, 2014.

[62] J. H. Ward Jr, Hierarchical grouping to optimize an objective function, Journal of the American statistical association 58 (301) (1963) 236–244.

[63] R. R. Sokal, F. J. Rohlf, The comparison of dendrograms by objective methods, Taxon 11 (2) (1962) 33–40.

[64] D. Defays, An efficient algorithm for a complete link method, The Computer Journal 20 (4) (1977) 364–366.

[65] F. Petitjean, A. Ketterlin, P. Gançarski, A global averaging method for dynamic time warping, with applications to clustering, Pattern recognition 44 (3) (2011) 678–693.

[66] F. Petitjean, G. Forestier, G. I. Webb, A. E. Nicholson, Y. Chen, E. Keogh, Dynamic time warping averaging of time series allows faster and more accurate classification, in: Data Mining (ICDM), 2014 IEEE International Conference on, IEEE, 2014, pp. 470–479.

[67] Y. Sakurai, M. Yoshikawa, C. Faloutsos, Ftw: fast similarity search under the time warping distance, in: Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, 2005, pp. 326–337.

[68] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979) 224–227. doi:10.1109/TPAMI.1979.4766909.

[69] S. Seabold, J. Perktold, statsmodels: Econometric and statistical modeling with python, in: 9th Python in Science Conference, 2010, pp. 92–96.

[70] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences 101 (suppl 1) (2004) 5228–5235.

[71] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.