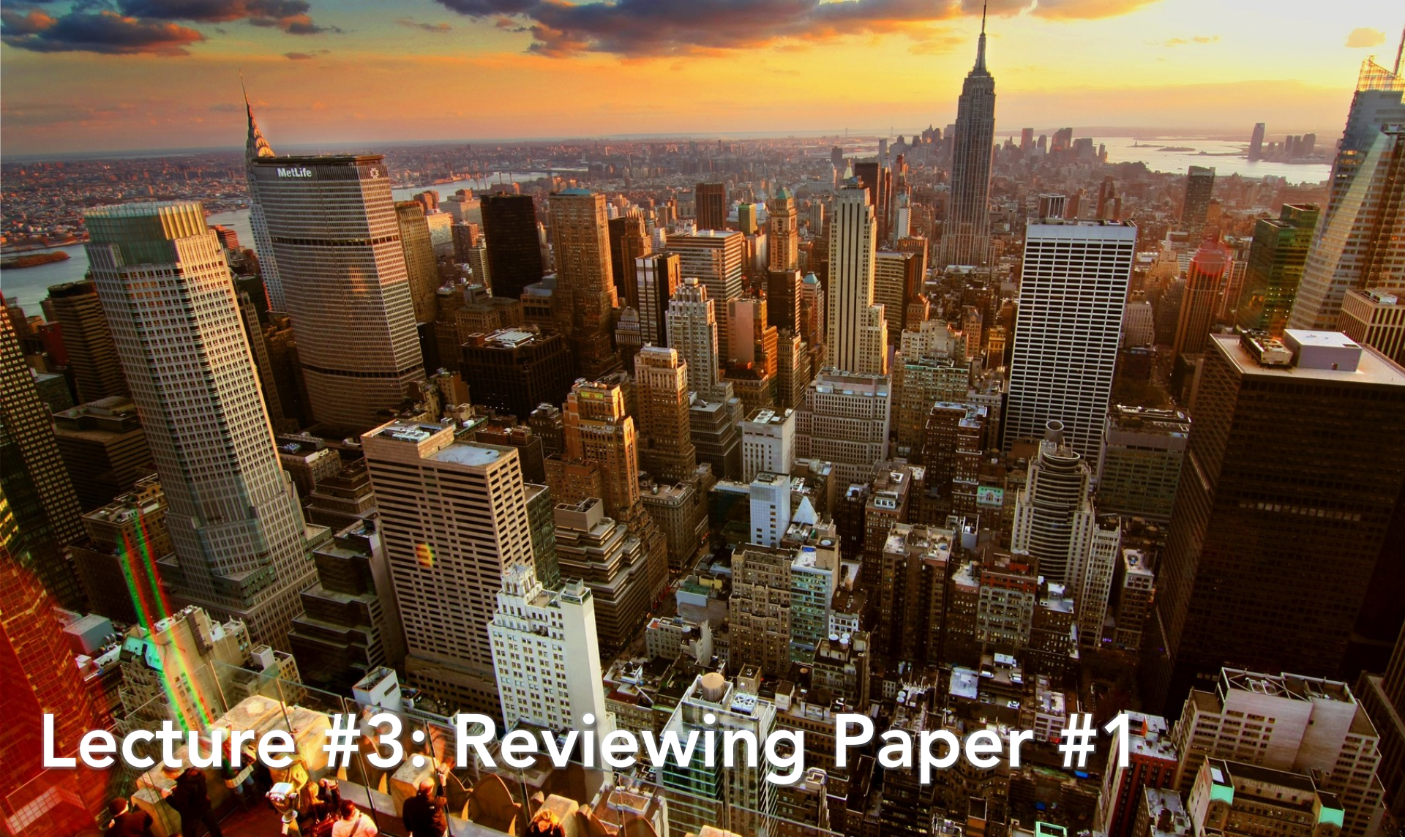


Lecture #3: Reviewing Paper #1

Mario Bergés
Associate Professor

12-752: Data-Driven Energy Management of Buildings

Civil & Environmental
ENGINEERING
Carnegie Mellon

Administrative stuff…

- Rescheduling classes.
- TA hours
- My office hours…
- Others?

Paper #1

Jungsuk Kwac, June Flora, Ram Rajagopal, **Household Energy Consumption Segmentation Using Hourly Data**, *IEEE Transactions on Smart Grid* , 2014

Household Energy Consumption Segmentation Using Hourly Data

Jungsuk Kwac, *Student Member, IEEE*, June Flora, and Ram Rajagopal, *Member, IEEE*

***Abstract*—The increasing US deployment of residential advanced metering infrastructure (AMI) has made hourly energy consumption data widely available. Using CA smart meter data, we investigate a household electricity segmentation methodology that uses an encoding system with a pre-processed load shape dictionary. Structured approaches using features derived from the encoded data drive five sample program and policy relevant energy lifestyle segmentation strategies. We also ensure that the methodologies developed scale to large data sets.**

***Index Terms*—Clustering, demand response, segmentation, smart meter data, variability.**

I. Introduction

THE WIDESPREAD deployment of advanced metering infrastructure (AMI) has made available concrete information about user consumption from smart meters. Household load shapes reveal significant differences among large groups of households in the magnitude and timing of their electricity consumption [3]. Hourly smart meter data offers a unique opportunity to understand a household's energy use lifestyle. Further, this consumption lifestyle information has the potential to enhance targeting and tailoring of demand response (DR) and energy efficiency (EE) programs as well as improving energy reduction recommendations. According to the Federal Energy Regulatory Commission, DR is defined as: "Changes in electric use by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized." EE means using less power to perform the same tasks, on a continuous basis or whenever that task is performed.

In this paper, an electricity customer segmentation methodology that uses an encoding system with a pre-processed load shape dictionary is examined. Energy consumers load shape information then is used to classify households according to extracted features such as entropy of shape code which measures the amount of *variability* in consumption. Load shape information enhances our ability to understand individual as well as groups of consumers. For example, time of day building occupancy and energy consuming activities can be interpreted from these shapes.

In the proposed segmentation system, we use a structured approach that uses features derived from the encoded data to drive the segmentation. We also develop segmentation strategies that aligned with specific application purposes such as household targeting for EE programs or recommendations for time of use shifts. In addition, we ensure that methods can readily scale to large data sets. We test our approach in a 220 K household data sample for a large utility.

A. Prior Work

Much of the previous research on audience segmentation takes place in psychology, marketing, and communication. Almost all segmentation in those fields rely on surveys of individuals regarding their self- reported values, attitudes, knowledge, and behaviors [1], [2]. In the last decade, utility companies are increasingly using these psychographic segmentation strategies to support program targeting, recruitment message tailoring and program design for DR and EE programs in [8]. Rarely, however is actual energy use part of the segmentation strategy [6]–[9].

Recently, the wide-spread dissemination of electricity smart meters offers the opportunity to create segmentation strategies based on 15 min, 30 min, or hourly household energy use. Understanding a households time of day energy consumption, daily usage pattern stability over time, as well as actual volume of energy use offers insights into household use of energy [3]. Further, these consumption features can be relevant to marketing and program design tasks. For example, high usage volume consumers or load shapes may signal potential for certain energy efficiency messages, whereas household load shape stability may be more relevant for time of use reduction messages.

Existing literature on analysis of smart meter data focuses on forecasting and load profiling such as [10]–[13], [18]. Some significant contributions in segmentation are [3], [4], [16], [17], [19], [20]. Self-organizing maps (SOM) and K-means are used to find load patterns in [17] and to present a electricity consumer characterization framework in [4]. A two-stage pattern recognition of load curves based on various clustering methods, including K-means, is described in [20]. Various clustering algorithms (hierarchical clustering, K-means, fuzzy K-means, SOM) are used to segment customers with similar consumption behavior in [16]. Similarly, [19] checks the capacity of SOM to filter, classify, and extract load patterns. As an alternative approach to distance-based clustering (K-means, SOM), [18] introduces a class of mixture models, random effects mixture

Manuscript received May 31, 2013; revised July 29, 2013; accepted July 29, 2013. Date of current version December 24, 2013. The work of R. Rajagopal is supported by the Powell Foundation Fellowship. Paper no. TSG-00427-2013.

J. Kwac is with the Department of Electrical Engineering and the Stanford Sustainable Systems Lab, Department of Civil and Environmental Engineering, Stanford University, CA 94305 USA (e-mail: kwjusu1@stanford.edu).

J. Flora is with the Human Sciences & Technologies Advanced Research Institute (HSTAR), Stanford University, CA 94305 USA (e-mail: jflora@stanford.edu).

R. Rajagopal is with the Stanford Sustainable Systems Lab, Department of Civil and Environmental Engineering, Stanford University, CA 94305 USA (e-mail: ramr@stanford.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSG.2013.2278477

Carnegie Mellon

Paper Discussion

- General thoughts

Review Process

Write Review

Offline reviewing Upload form: Choose File No file chosen Go

Download form · **Tip:** Use Search or Offline reviewing to download or upload many forms at once.

Overall merit

(Your choice here)

Reviewer expertise

(Your choice here)

Paper summary

Markdown styling and LaTeX math supported

Comments for author

Markdown styling and LaTeX math supported

Comments for PC (hidden from authors)

Markdown styling and LaTeX math supported

Paper Discussion

- What is the general problem being addressed?
- What are the novel scientific/engineering contributions?
- What was the approach taken (factual)?
  - Is the approach validated?
- What are the key findings?
  - Are the results validated?

Paper Discussion

- Smart Meter Datasets
  - http://wiki.nilm.eu/index.php?title=NILM_datasets
- Tools:
  - Exploratory Data Analysis
  - Clustering

Paper figures/tables

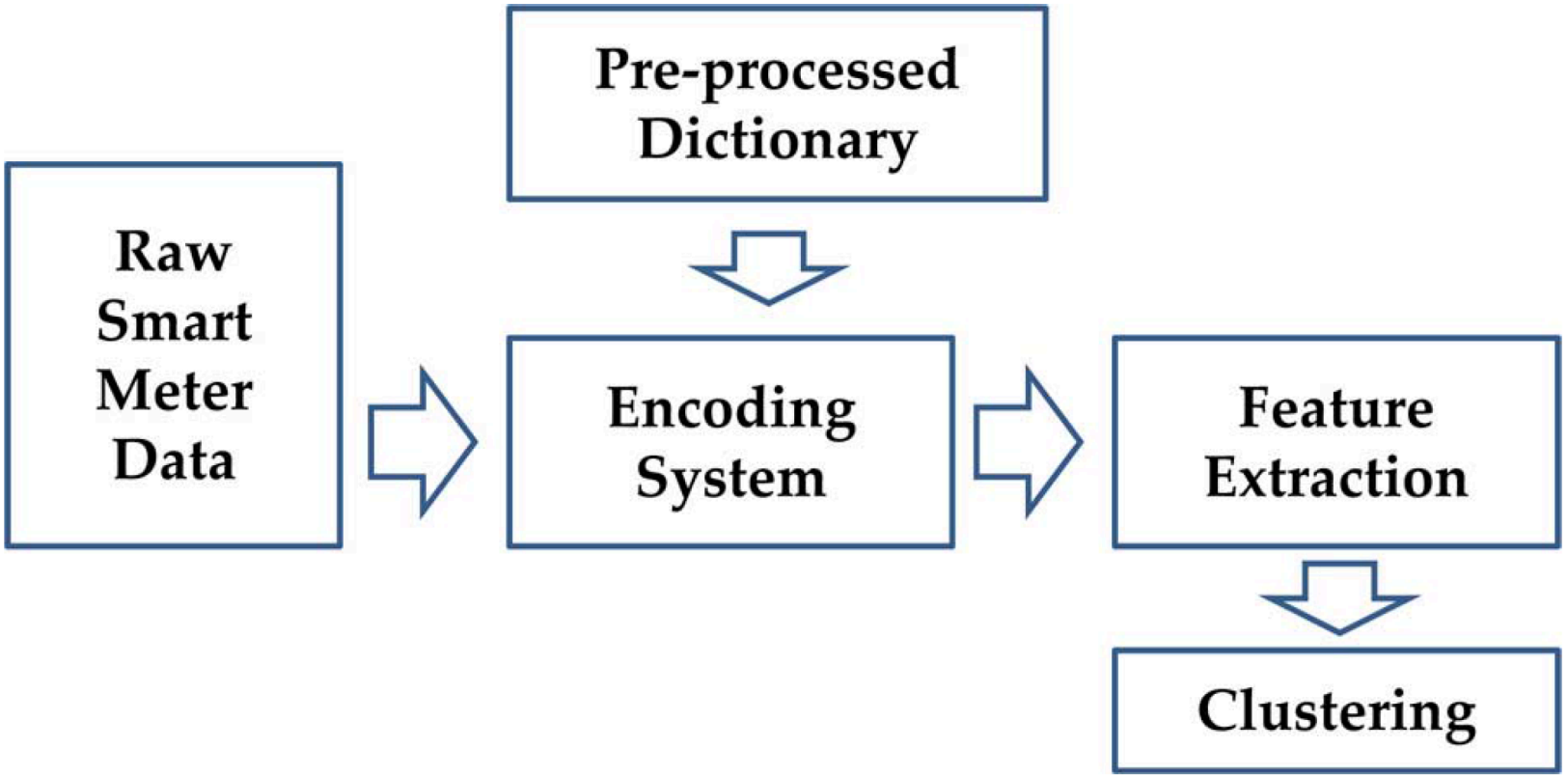


Fig. 1. User segmentation flow.

Equation 1:

$$a = \sum_{t=1}^{24} l(t) \text{ and } s(t) = \frac{l(t)}{a}. \quad (1)$$

Paper figures/tables

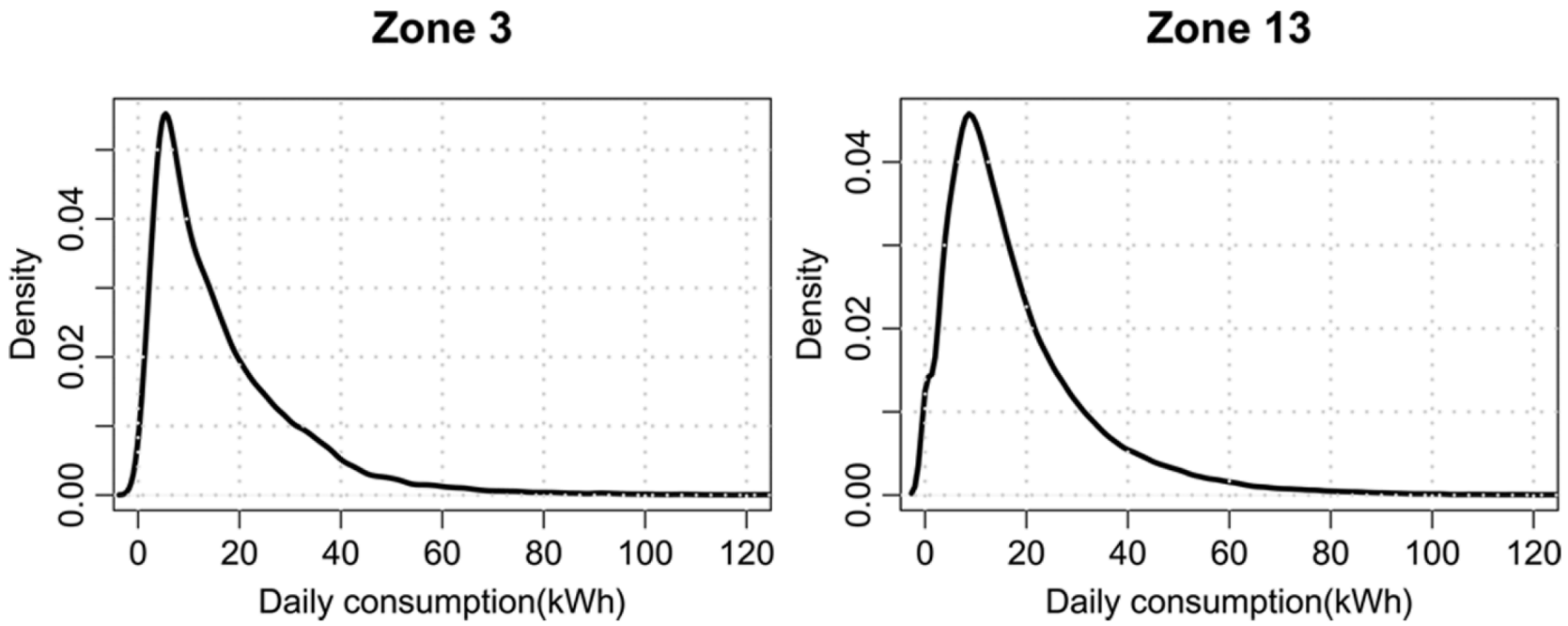


Fig. 2. Daily consumption distribution at Zone 3 and Zone 13.

Equation 2:

$$f(a) = \sum_{i=1}^{n} \lambda_i g_i(a), g_i(a) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{\frac{-(\log(1+a)-\mu_i)^2}{2\sigma_i^2}}, \quad (2)$$

Paper figures/tables

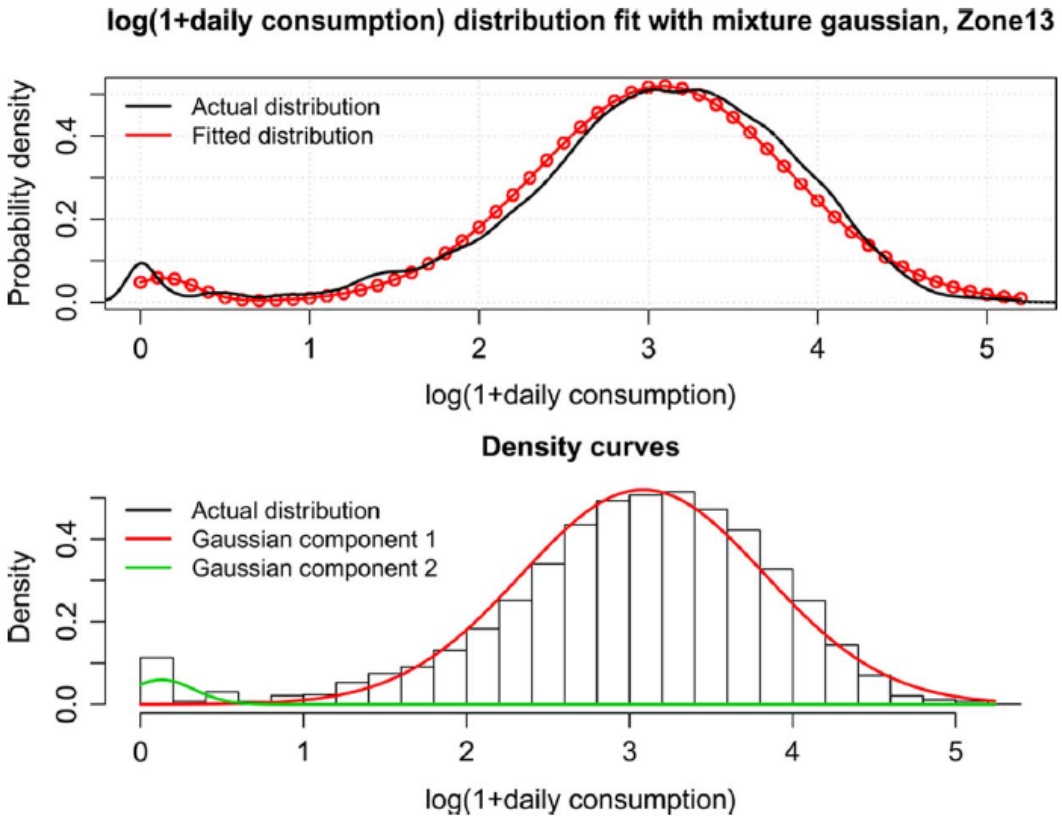


Fig. 3. Mixture of log normal distribution fitting on one zip code area for 2011 June–Aug. data.

Carnegie Mellon

Equation 3:

$$E(s,i) = \sum_{t=1}^{24} (C_i(t) - s(t))^2,$$
$$i^*(s) = \arg\min_i E(s,i). \quad (3)$$

Equation (4):

$$E(s, i^*(s)) = \sum_{t=1}^{24} \left(s(t) - C_{i^*(s)}(t)\right)^2 \leq \theta \sum_{t=1}^{24} C_{i^*(s)}(t)^2, \tag{4}$$

Clusters vs. Threshold

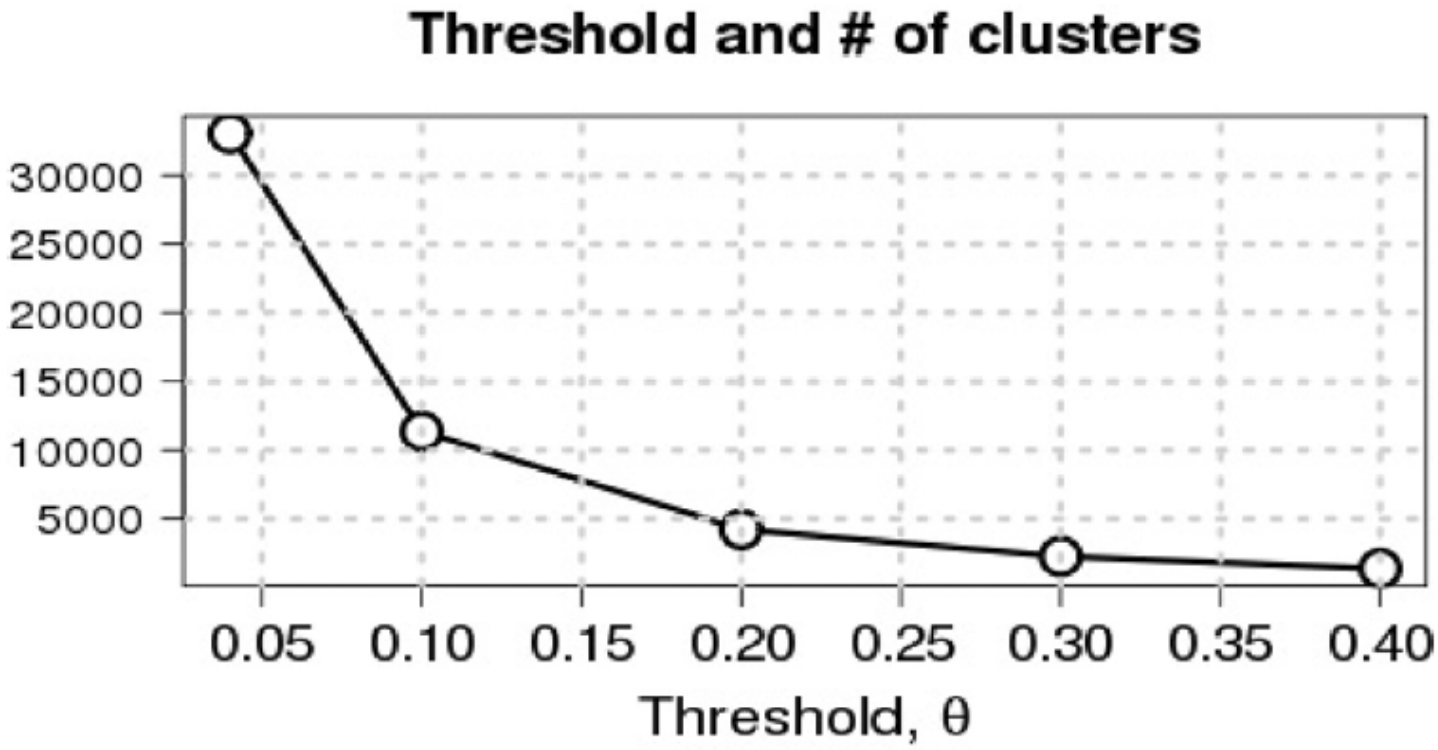


Fig. 4. Relation between threshold choice and number of clusters applying Algorithm 1.

Clustering Results

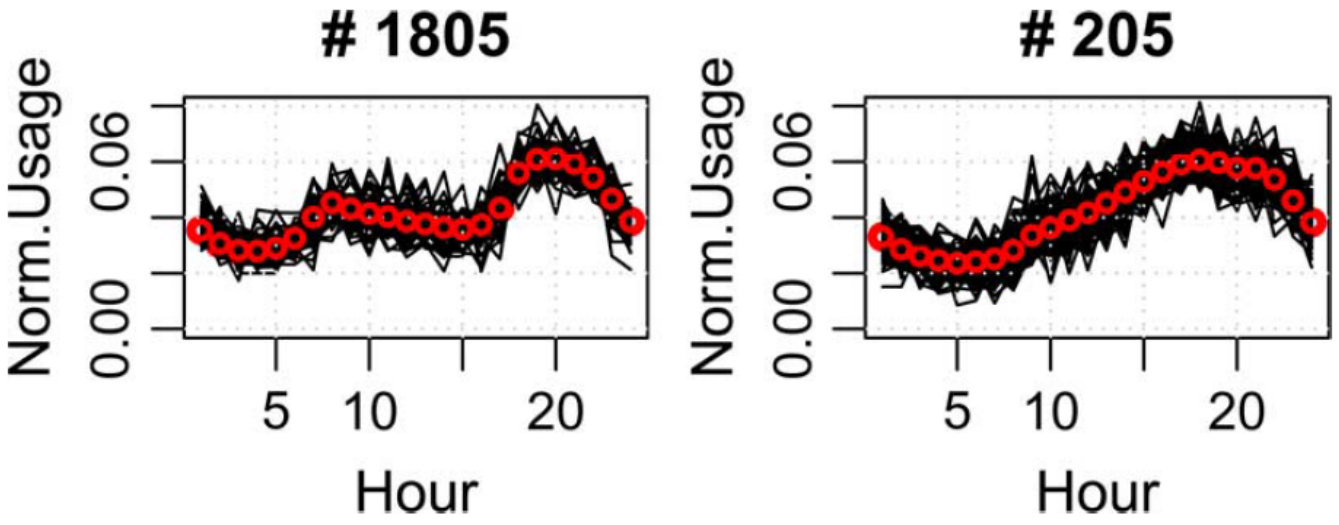


Fig. 5. Example of adaptive K-means result with $\theta = 0.2$: Normalized daily usage patterns (load shapes) and cluster centers.

Clustering Results

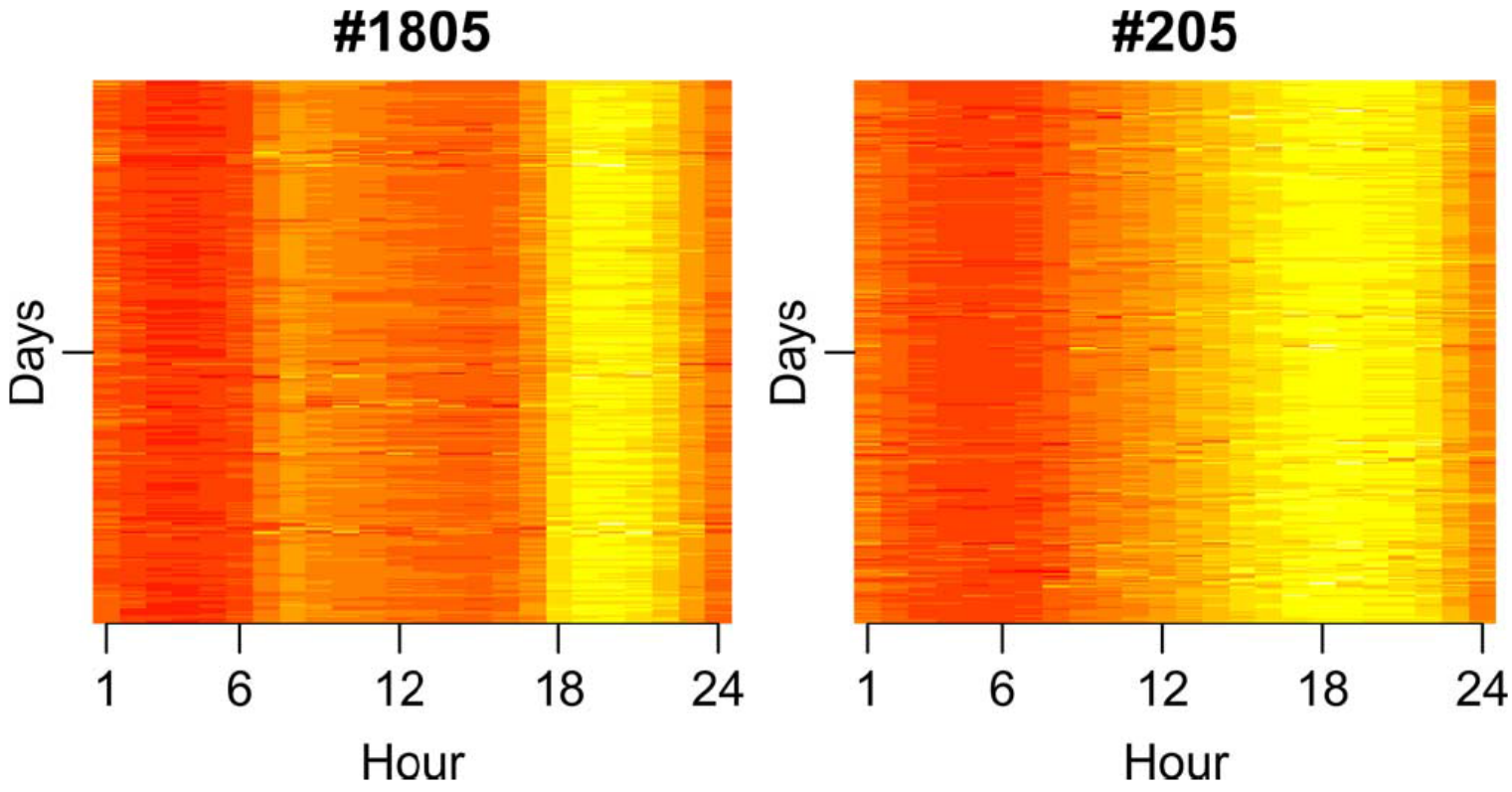


Fig. 6. Example of adaptive K-means result with $\theta = 0.2$: Heat map of normalized data under the same shape code.

Dictionary Coverage

TABLE I
DICTIONARY COVERAGE FROM ZONE 13

Zone 13	Zone 3 coverage	Zone 2 coverage
A dictionary with $\theta = 0.2$ populated	141876 (/143915) load shapes (98.6%)	85393 (/88771) load shapes (96.2%)

Cluster correlation and size

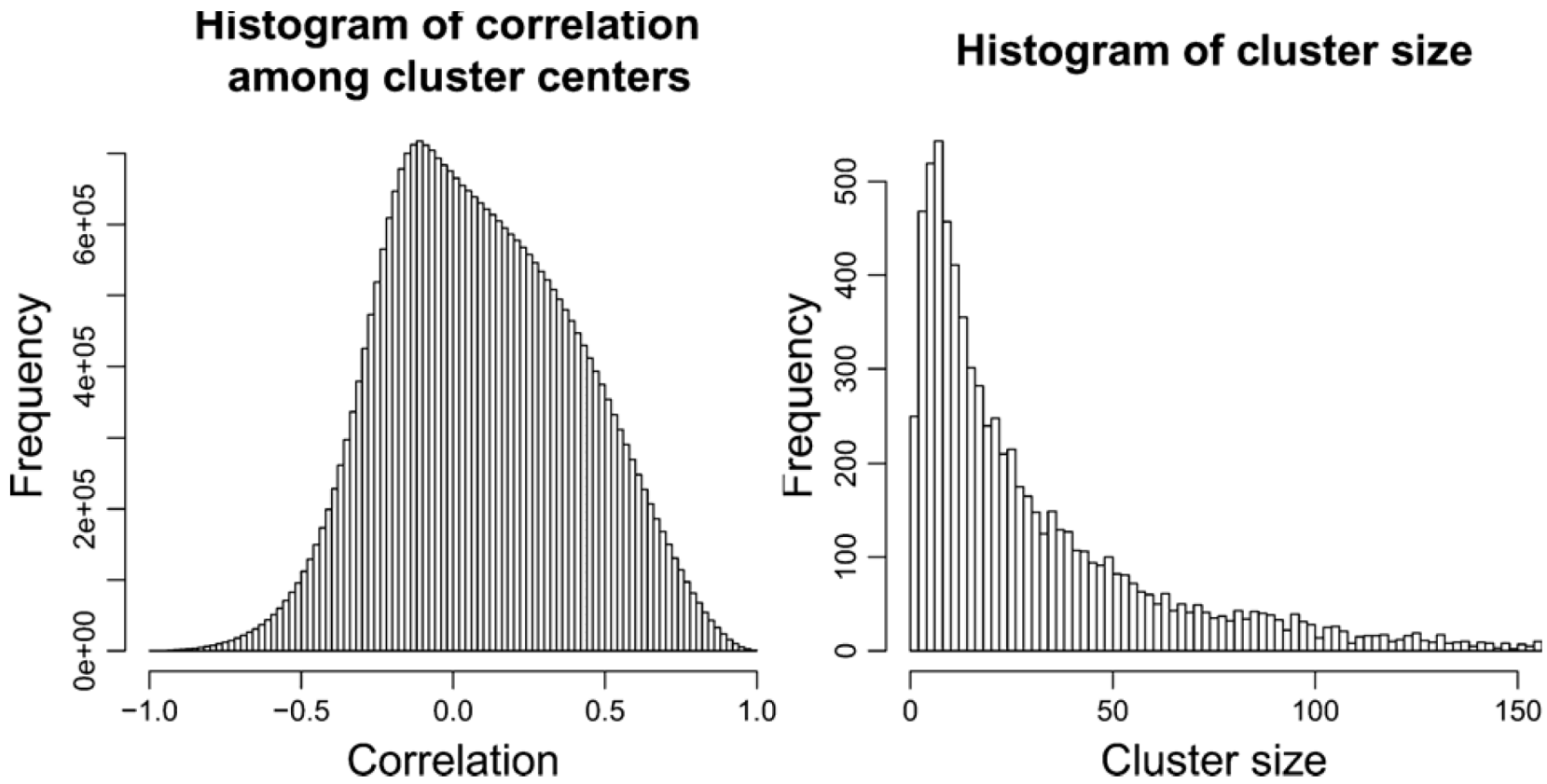


Fig. 7. Clusters correlation and size distribution.

Equation 5

$$\hat{\theta} = \frac{\sum_{t=1}^{24} (s(t) - C_{i^*(s)}(t))^2}{\sum_{t=1}^{24} \left(C_{i^*(s)}(t)\right)^2}, \qquad (5)$$

Figure 8

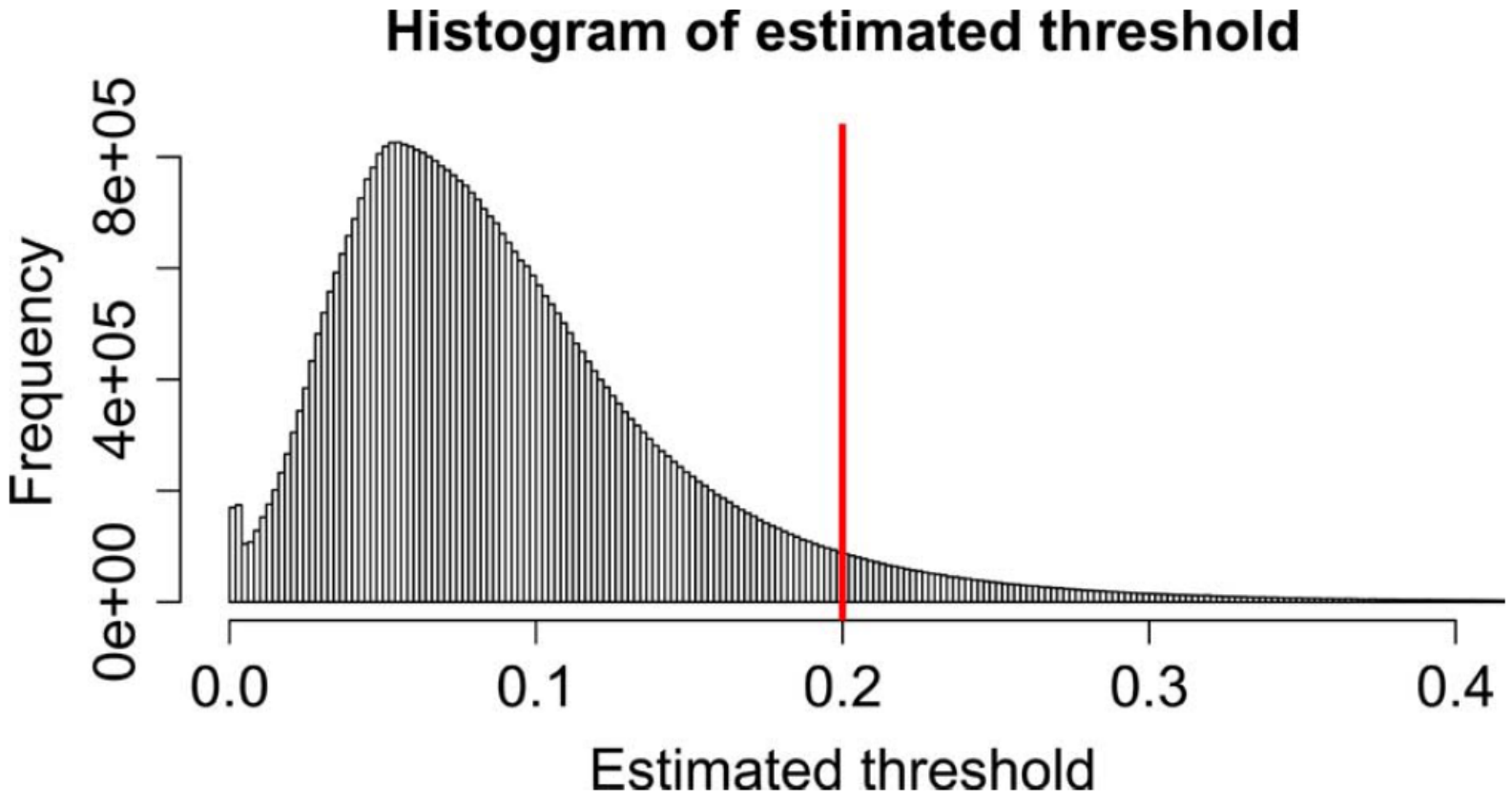


Fig. 8. Estimated threshold distribution.

Figure 8

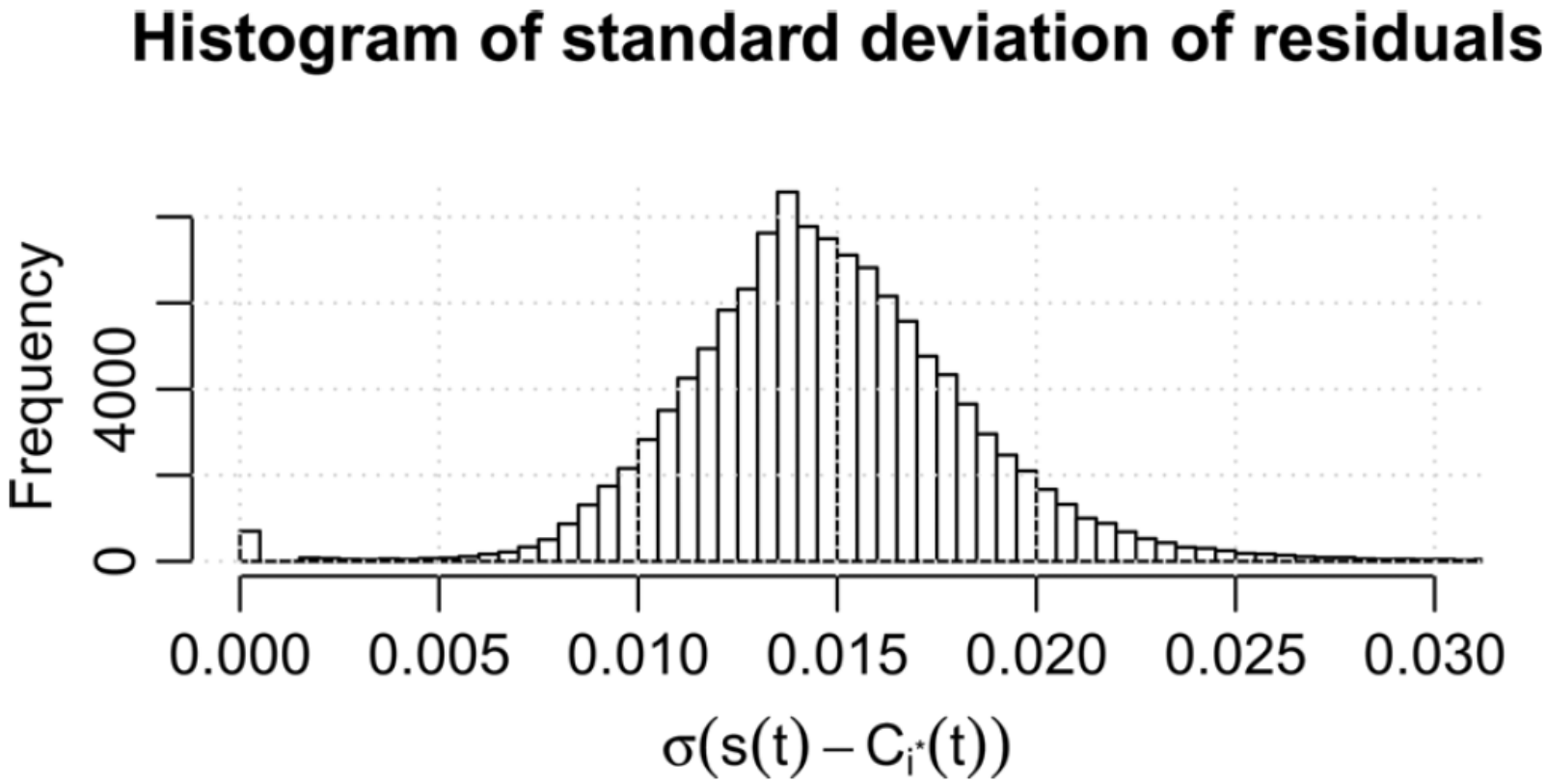


Fig. 9. Distribution of $\sigma(s(t) - C_{i^*(s)}(t))$.

Figure 10

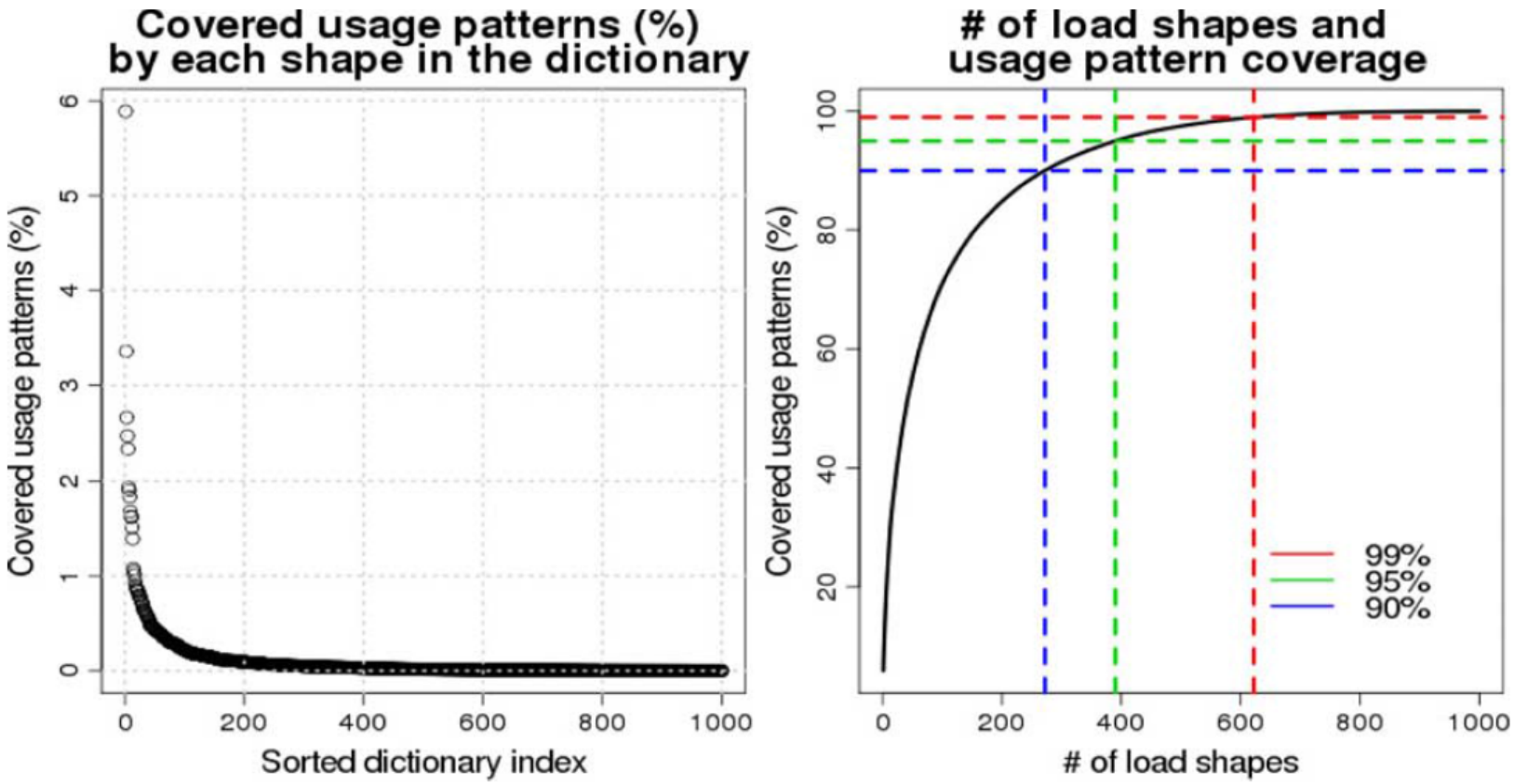


Fig. 10. Covered usage patterns and # of load shapes.

Load Shapes

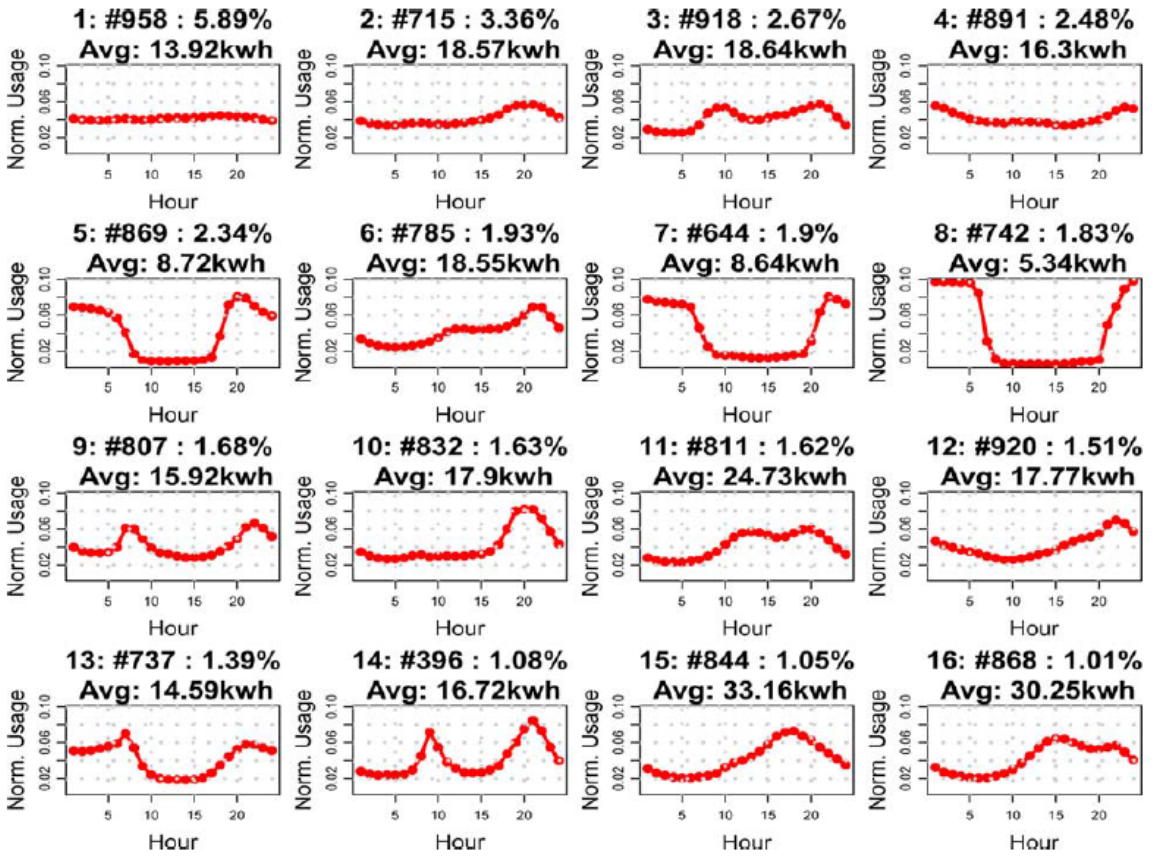


Fig. 11. 16 Most frequent load shapes of whole households.

Carnegie Mellon

Load Shape Popularity

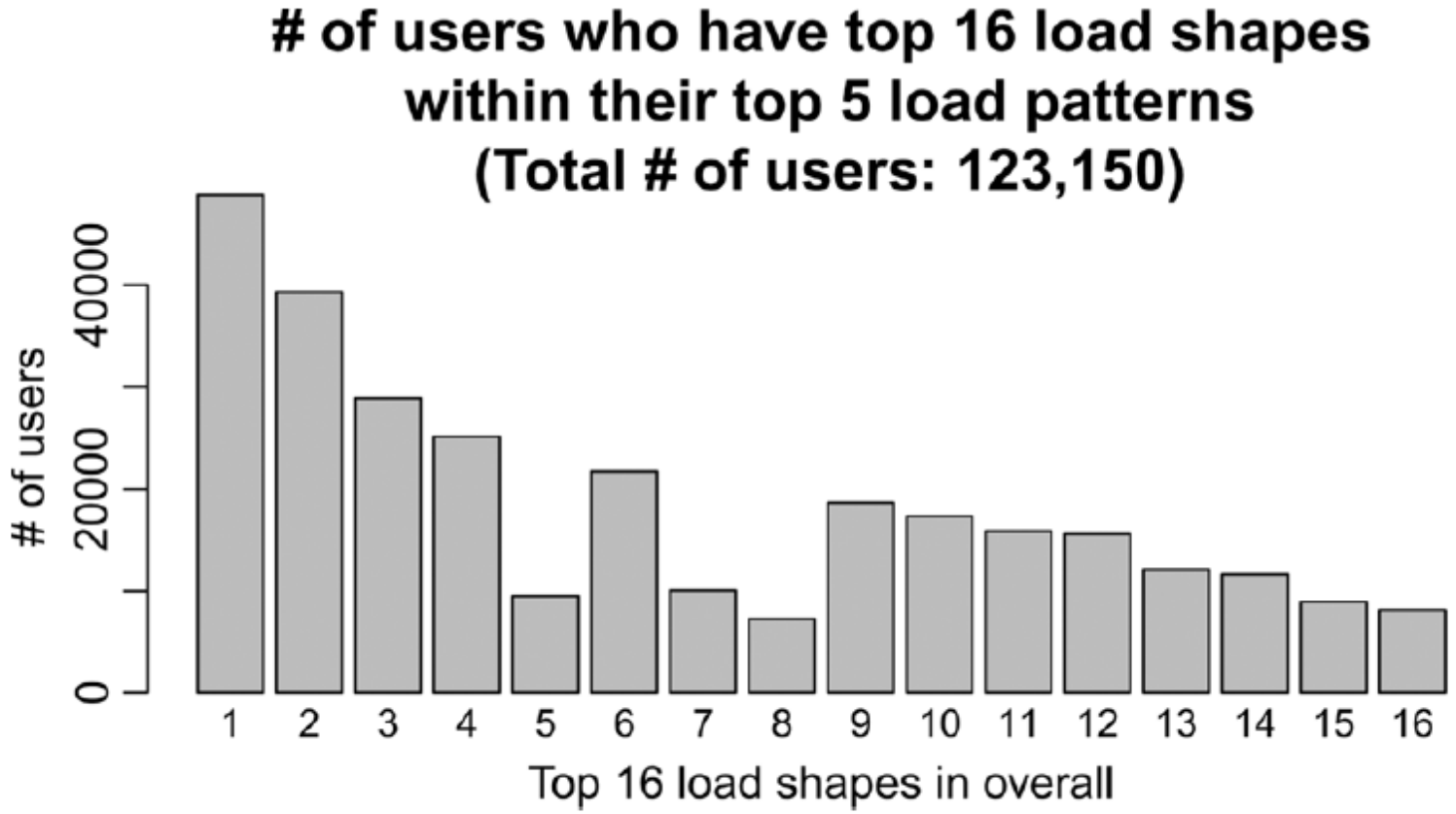


Fig. 12. # of users who have top 16 load shapes within their top 5 load patterns.

Load Shape Distribution

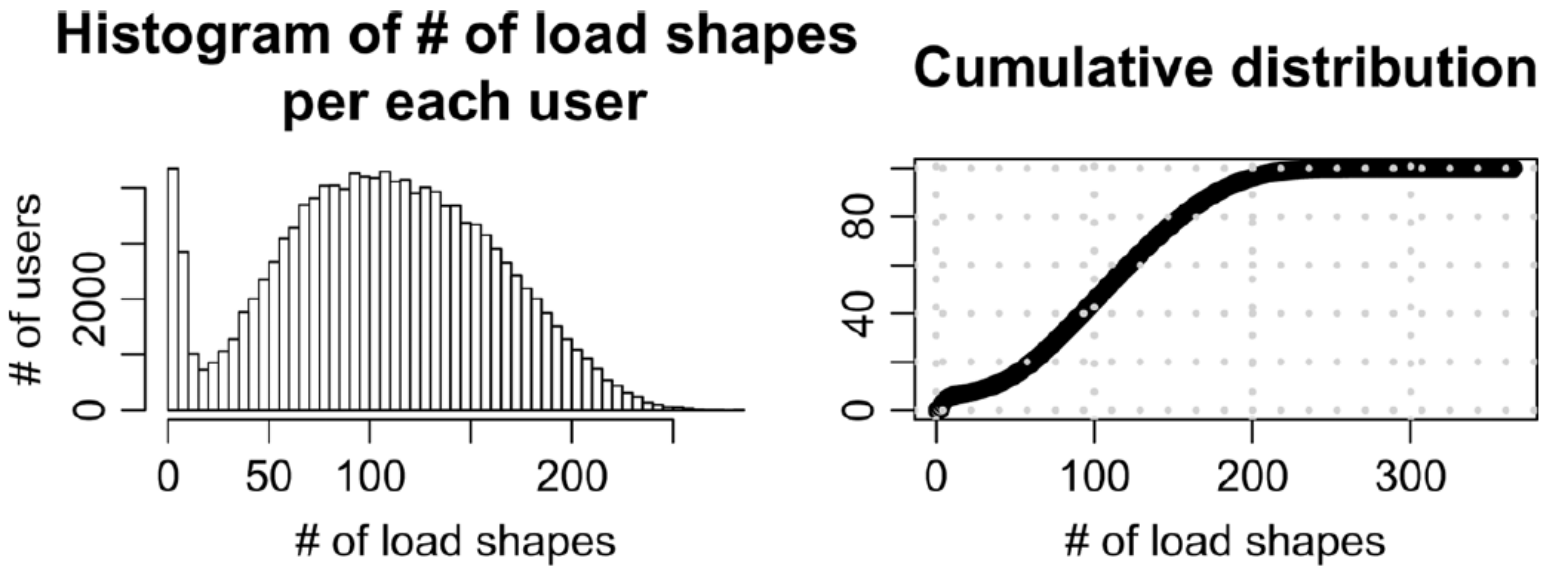


Fig. 13. # of load shapes per each user.

Equation 6

$$S_n = -\sum_{i=1}^{K} p(C_i) \log p(C_i). \quad (6)$$

Entropy Distribution

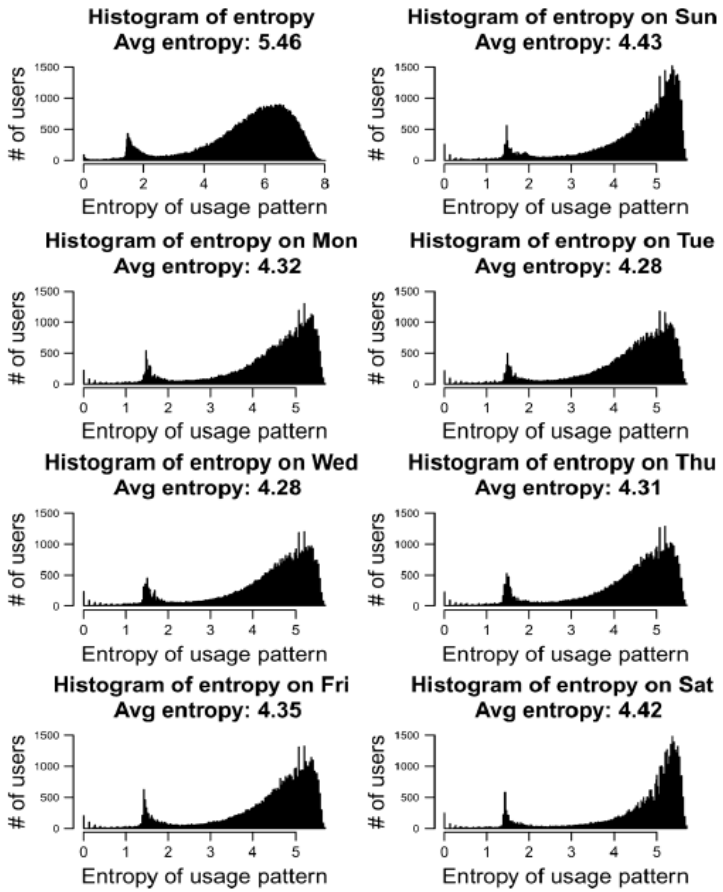
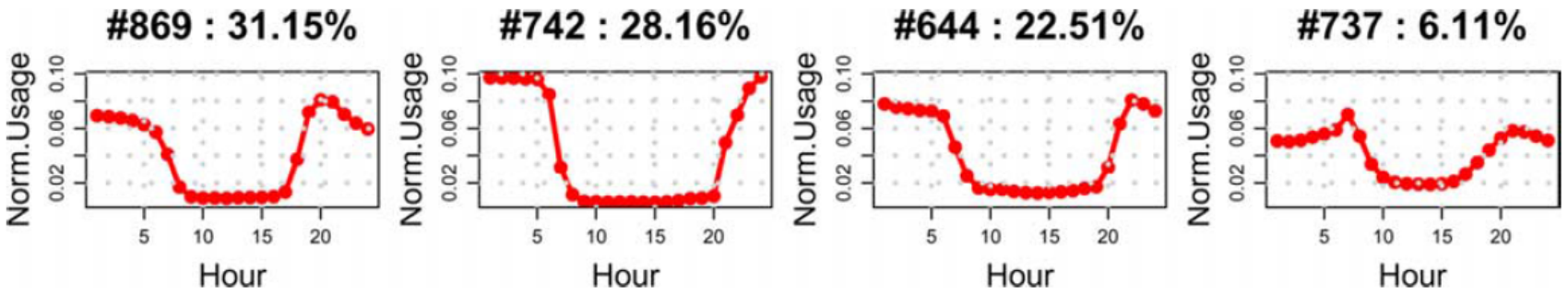


Fig. 14. Load shape entropy distribution.

Carnegie Mellon

Most Frequent Load Shapes in Low-Entropy Group



Carnegie Mellon

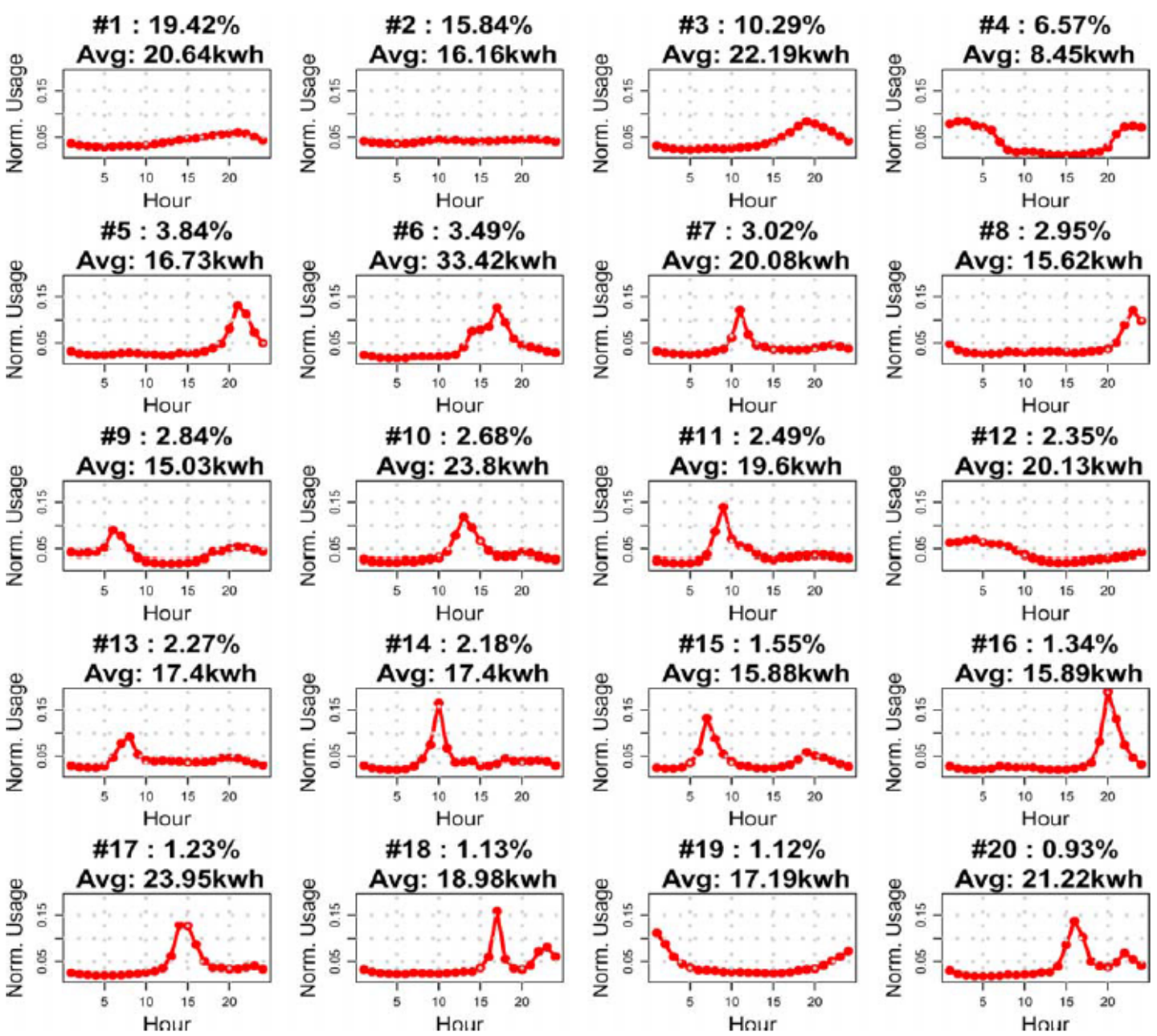


Fig. 16. 20 most frequent load shapes of whole households using the dictionary of size $= 100$.

Title = Caption

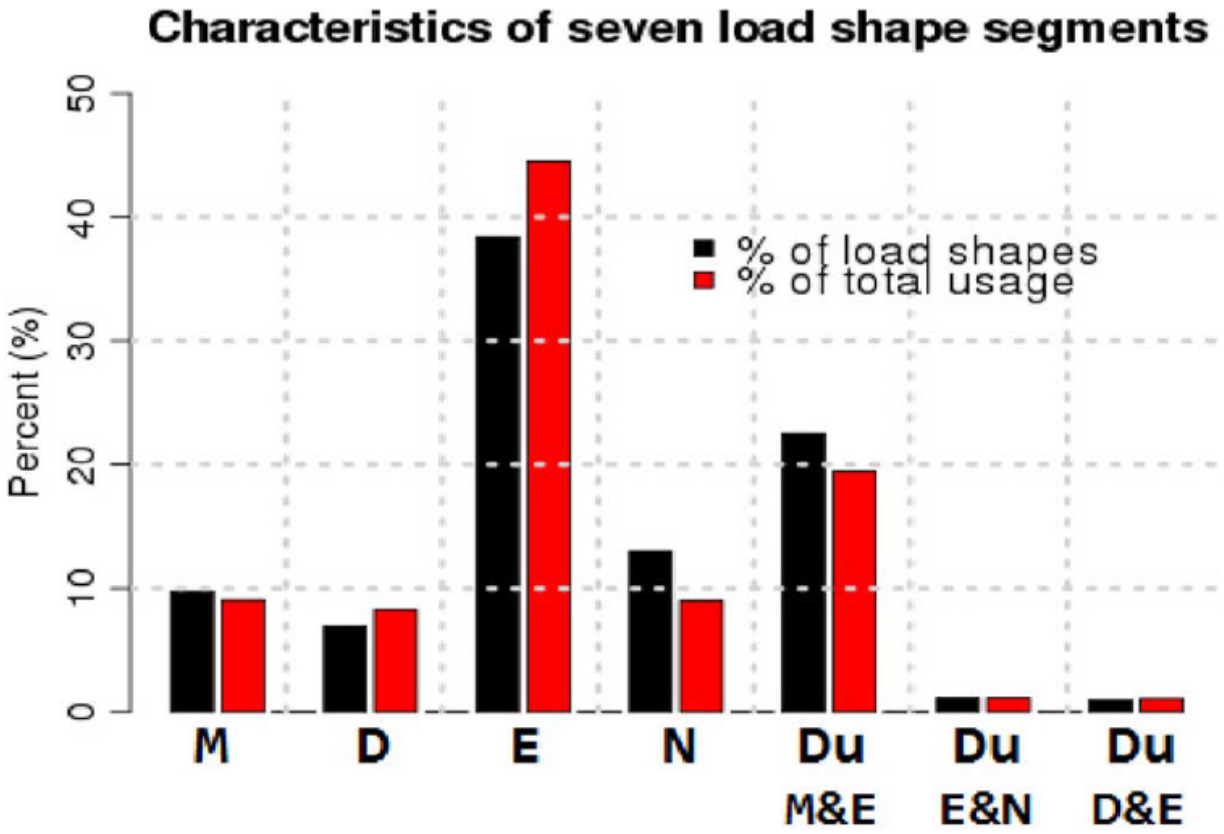


Fig. 17. Characteristics of seven load shape segments.

Segments per user

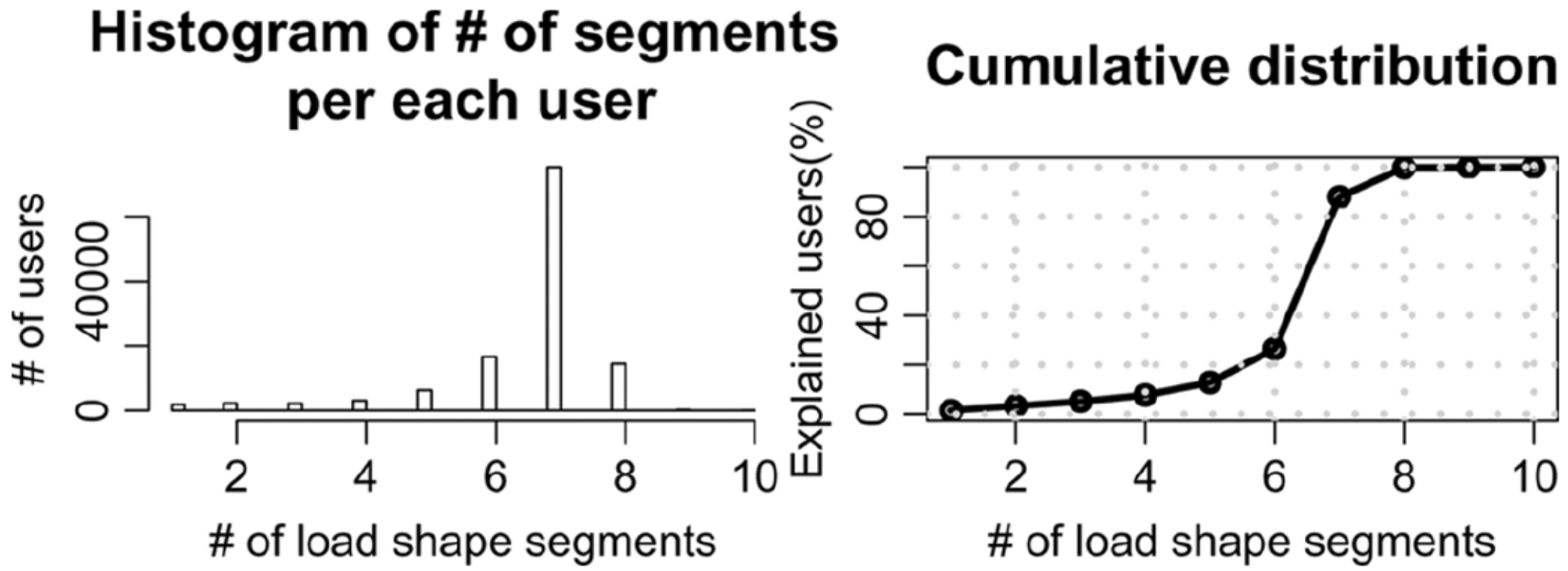


Fig. 18. # of load shape segments per each user.

Figure 19

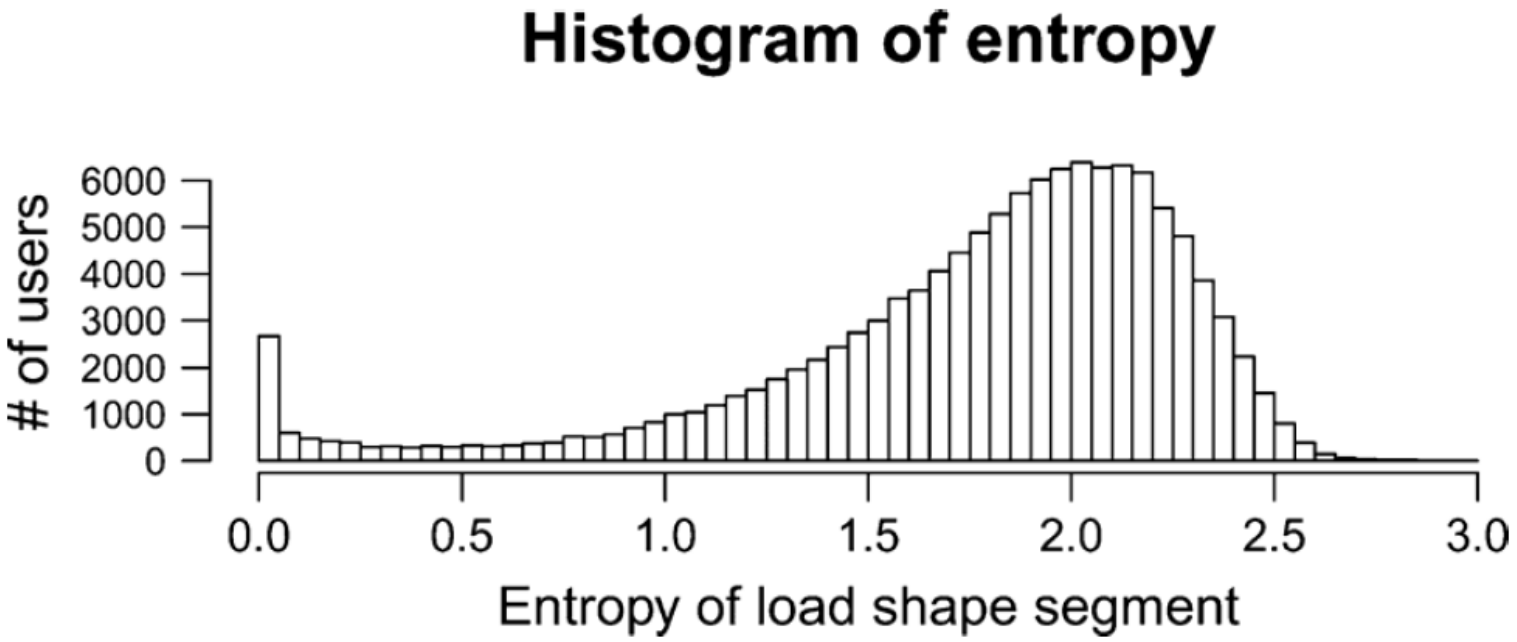


Fig. 19. Load shape segment entropy distribution.

Figure 20

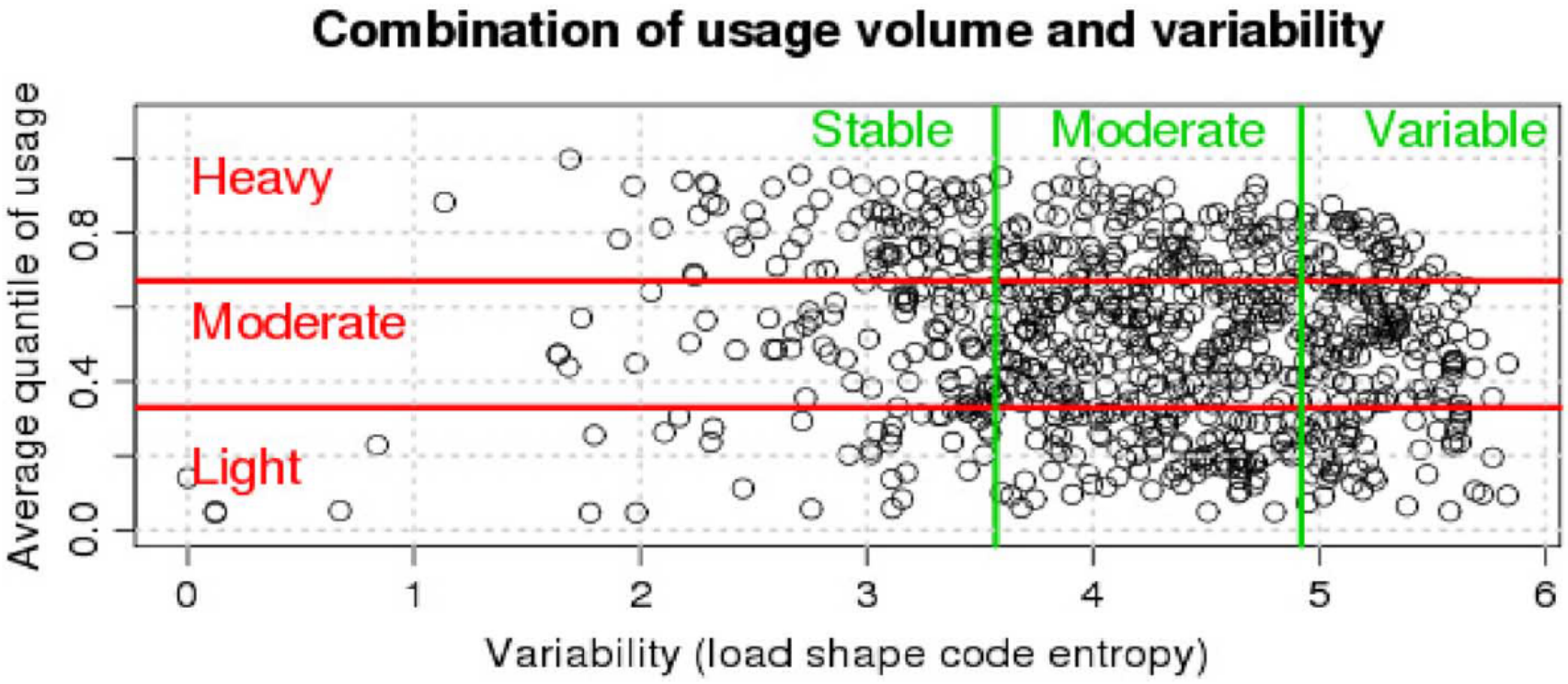


Fig. 20. Combination of usage volume and shape variability.

Table II

TABLE II
THE NUMBER OF HOUSEHOLDS IN FIG. 20

	Stable	Moderate	Variable	Total
Heavy	79 (10.2%)	103 (13.2%)	38 (4.9%)	220 (28.3%)
Moderate	73 (9.4%)	187 (24.1%)	106 (13.6%)	366 (47.1%)
Light	40 (5.1%)	100 (12.9%)	51 (6.6%)	191 (24.6%)
Total	192 (24.7%)	390 (50.2%)	195 (25.1%)	777 (100%)

Figure 21

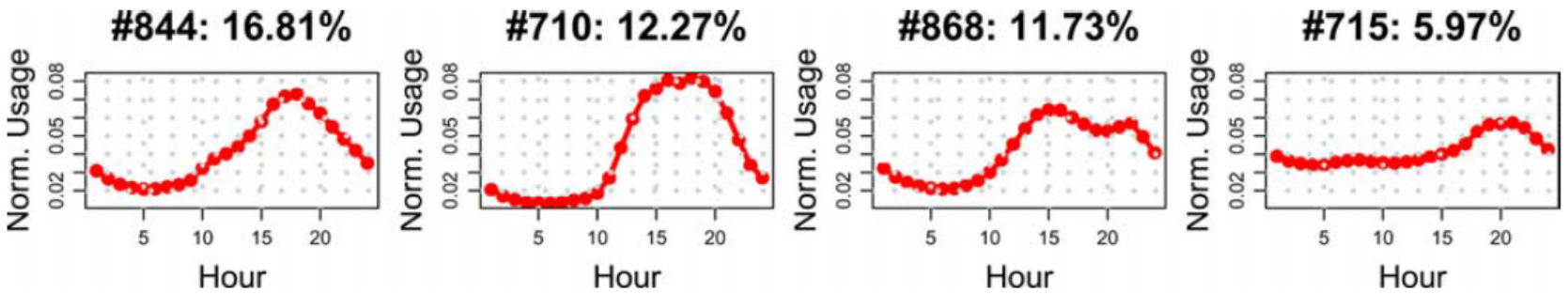


Fig. 21. Most frequent load shapes in the filtered users.

Table III

TABLE III
COMPARING LOAD SHAPE FREQUENCIES AMONG GROUPS

t-test to check $P(C_i|condition\ A) = P(C_i|condition\ B)$

N_1: sample size satisfying condition A

N_2: sample size satisfying condition B

$\bar{X_1}$: $\frac{\#\ of\ C_i\ among\ N_1}{N_1}$, $\bar{X_2}$: $\frac{\#\ of\ C_i\ among\ N_2}{N_2}$

$S_1^2 = \bar{X_1}(1-\bar{X_1})$, $S_2^2 = \bar{X_2}(1-\bar{X_2})$

$$T = \frac{\bar{X_1} - \bar{X_2}}{\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}\sqrt{\frac{(N_1-1)S_1^2+(N_2-1)S_2^2}{N_1+N_2-2}}}, \quad d.f. = N_1 + N_2 - 2$$

1) $T < t_{0.025}$: $P(C_i|condition\ A) < P(C_i|condition\ B)$
2) $T > t_{0.975}$: $P(C_i|condition\ A) > P(C_i|condition\ B)$
3) Otherwise: $P(C_i|condition\ A) = P(C_i|condition\ B)$

Zone 3

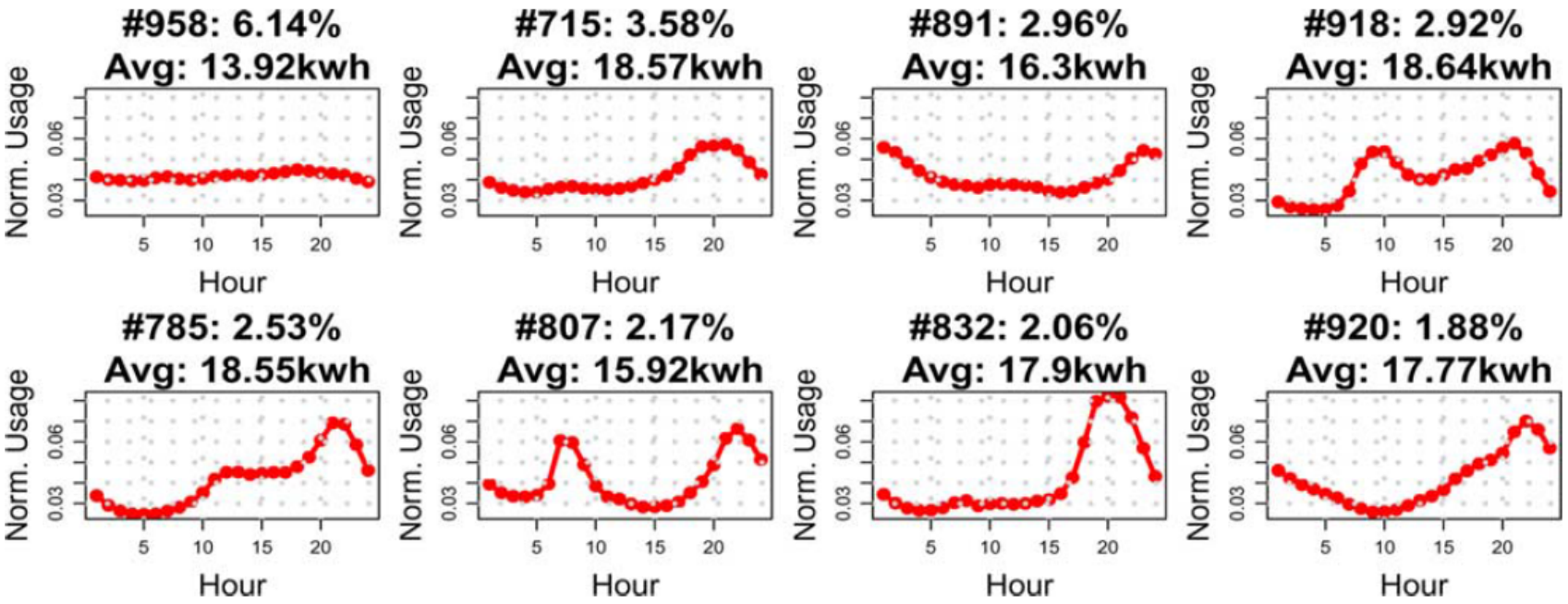


Fig. 22. More frequent load shapes in Zone 3.

Zone 12

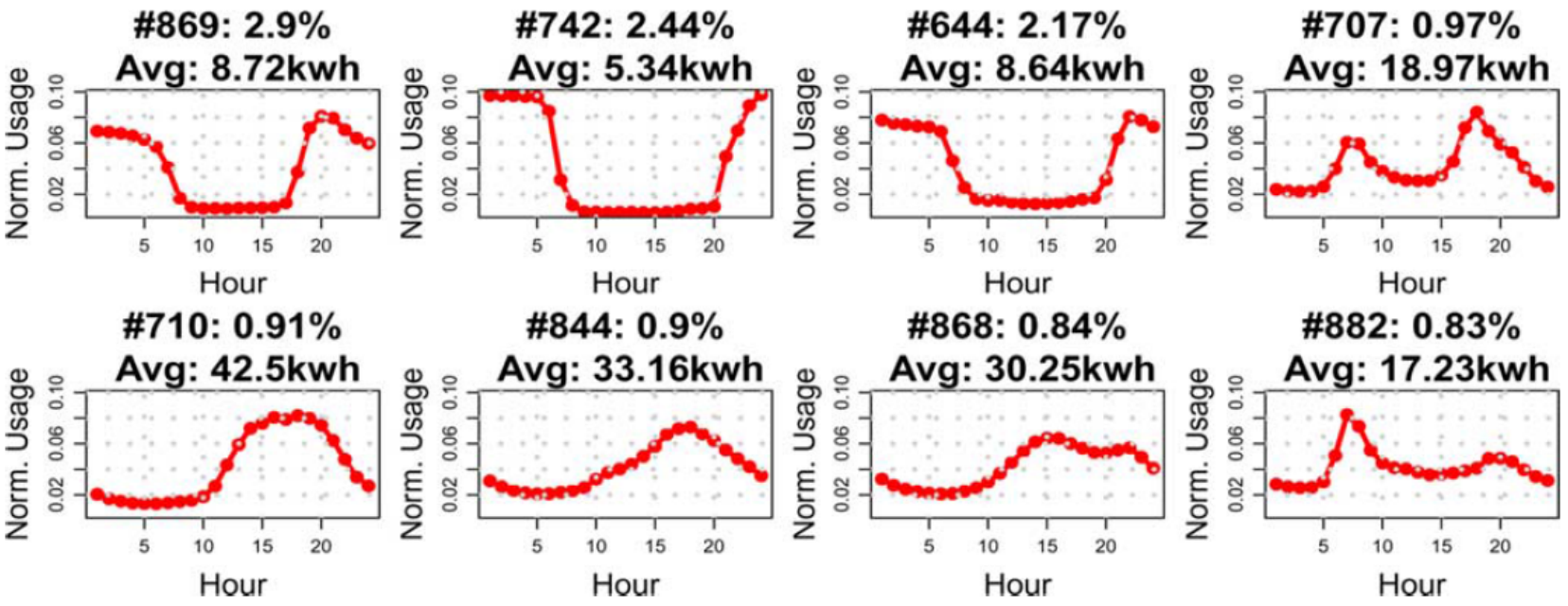


Fig. 23. More frequent load shapes in Zone 12.

Zones 3 and 12

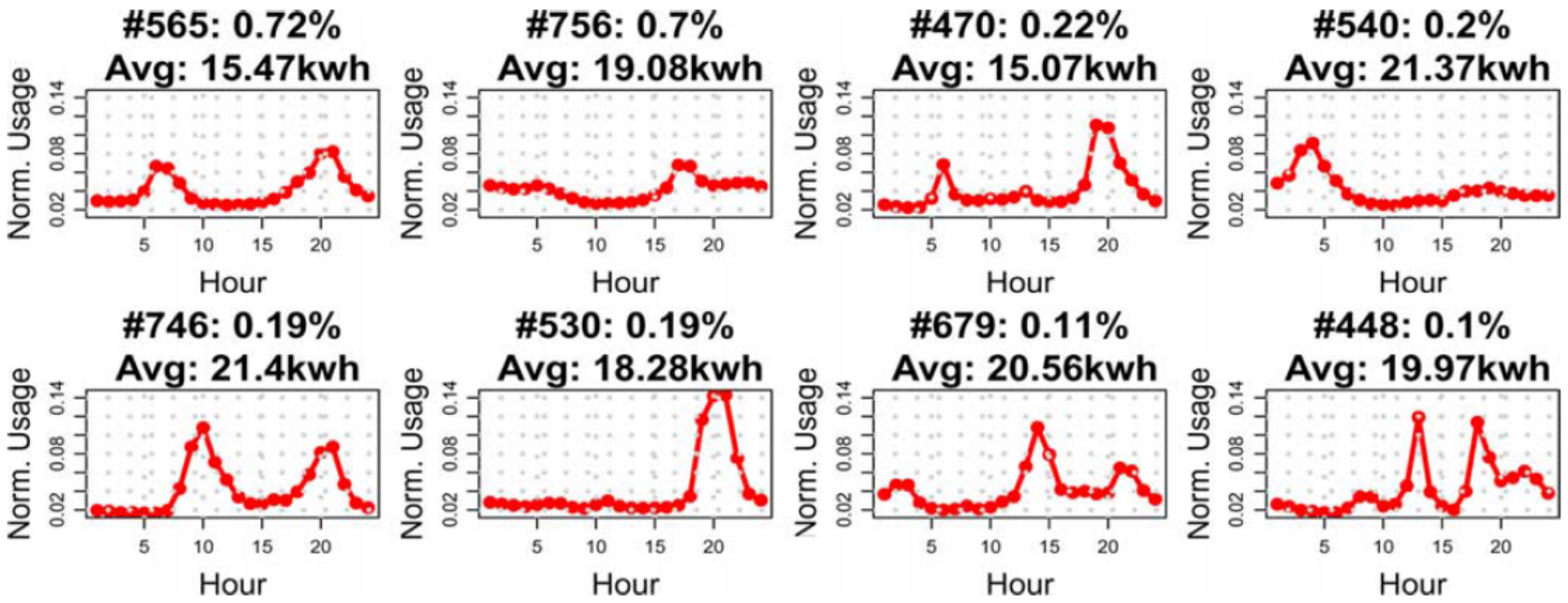


Fig. 24. Common load shapes in both zones.

Weekdays

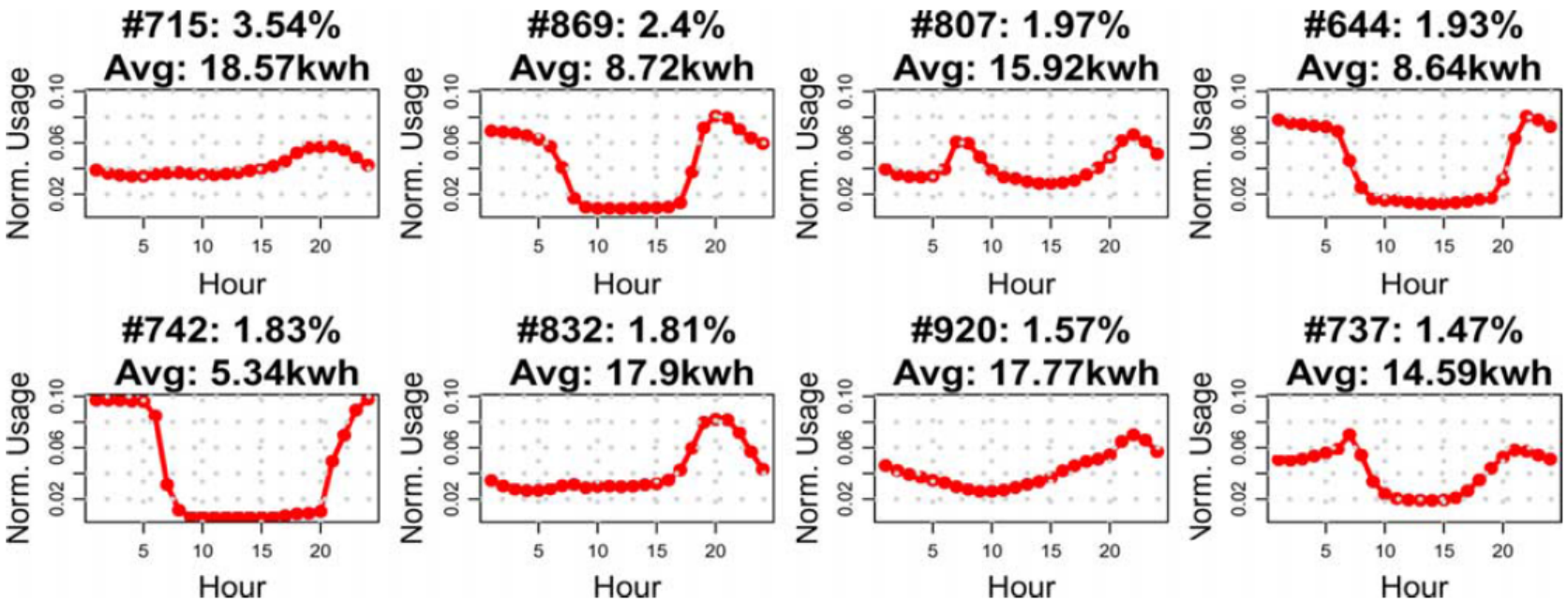


Fig. 25. More frequent load shapes in Weekdays.

Weekends

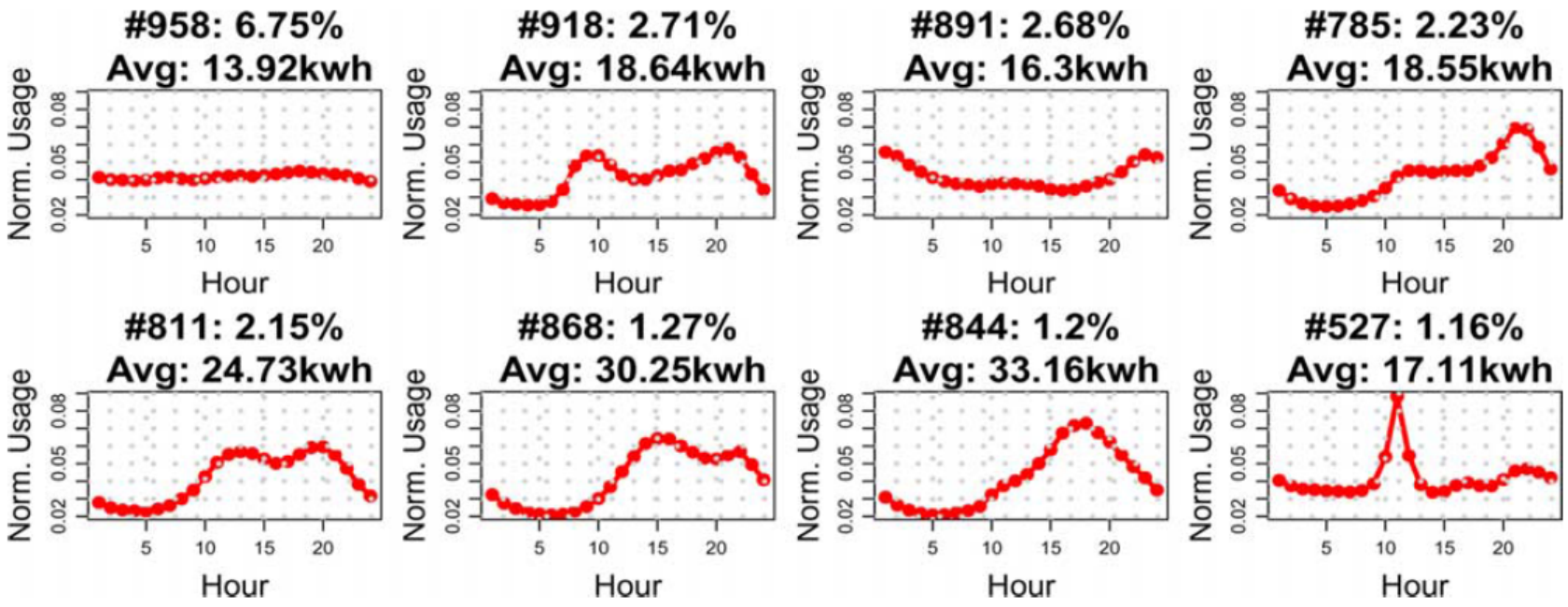


Fig. 26. More frequent load shapes in Weekends.

Weekends and Weekdays

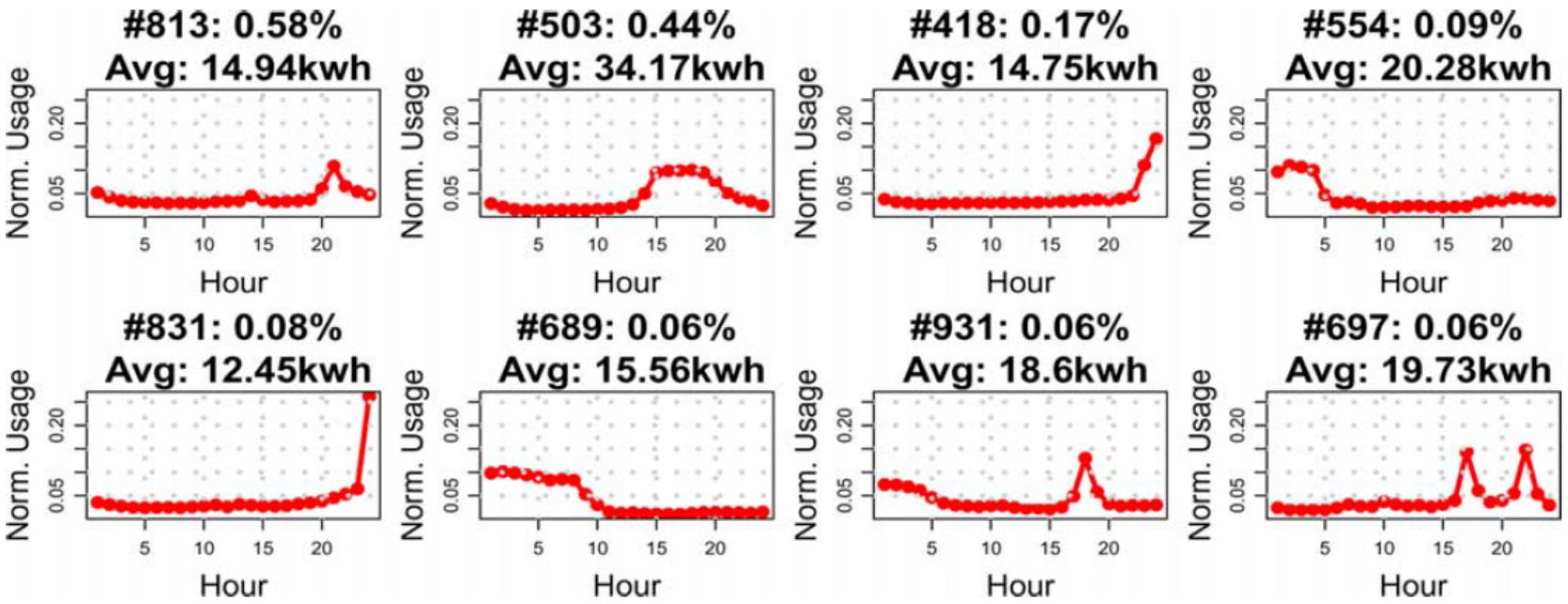


Fig. 27. Common load shapes in Weekdays & Weekends.

T-Test Results

TABLE IV
T-TEST RESULT

$P(C_i\|Zone3) > P(C_i\|Zone12)$	272
$P(C_i\|Zone3) < P(C_i\|Zone12)$	532
$P(C_i\|Zone3) = P(C_i\|Zone12)$	196
Total	1000

TABLE V
T-TEST RESULT

$P(C_i\|Weekdays) > P(C_i\|Weekends)$	322
$P(C_i\|Weekdays) < P(C_i\|Weekends)$	496
$P(C_i\|Weekdays) = P(C_i\|Weekends)$	182
Total	1000

A SLIGHT DETOUR…

A SLIGHT DETOUR…

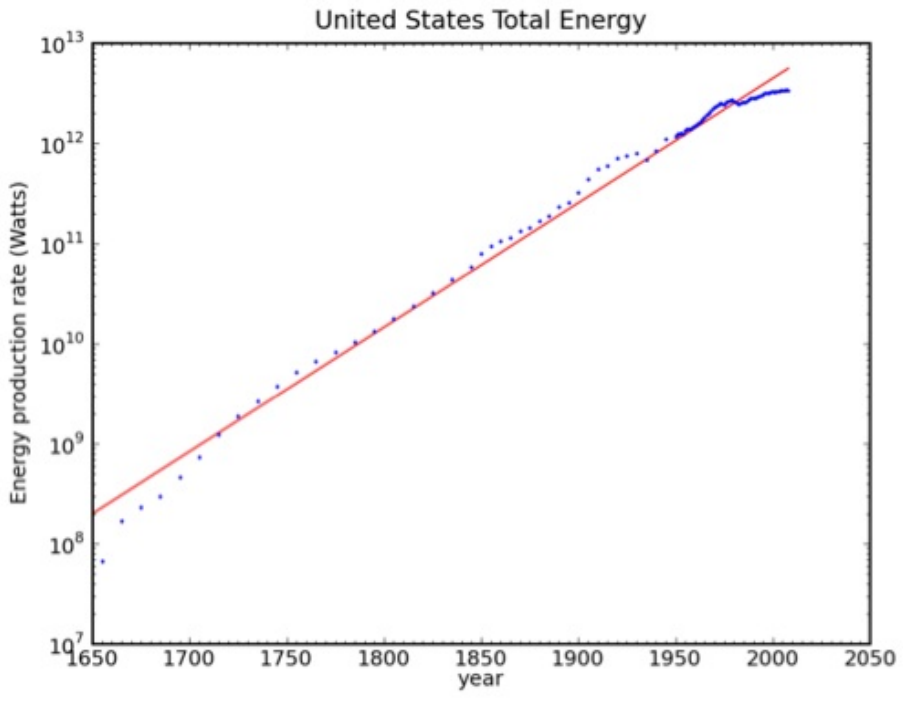
An Exercise in Galactic-Scale Energy

- Tom Murphy, UCSD

  – Is energy demand growth sustainable?

Carnegie Mellon

Energy Demand Growth



— Total U.S. Energy consumption in all forms since 1650. The vertical scale is logarithmic, so that an exponential curve resulting from a constant growth rate appears as a straight line. The red line corresponds to an annual growth rate of 2.9%. Data source: EIA.

Source: Tom Murphy

Questions

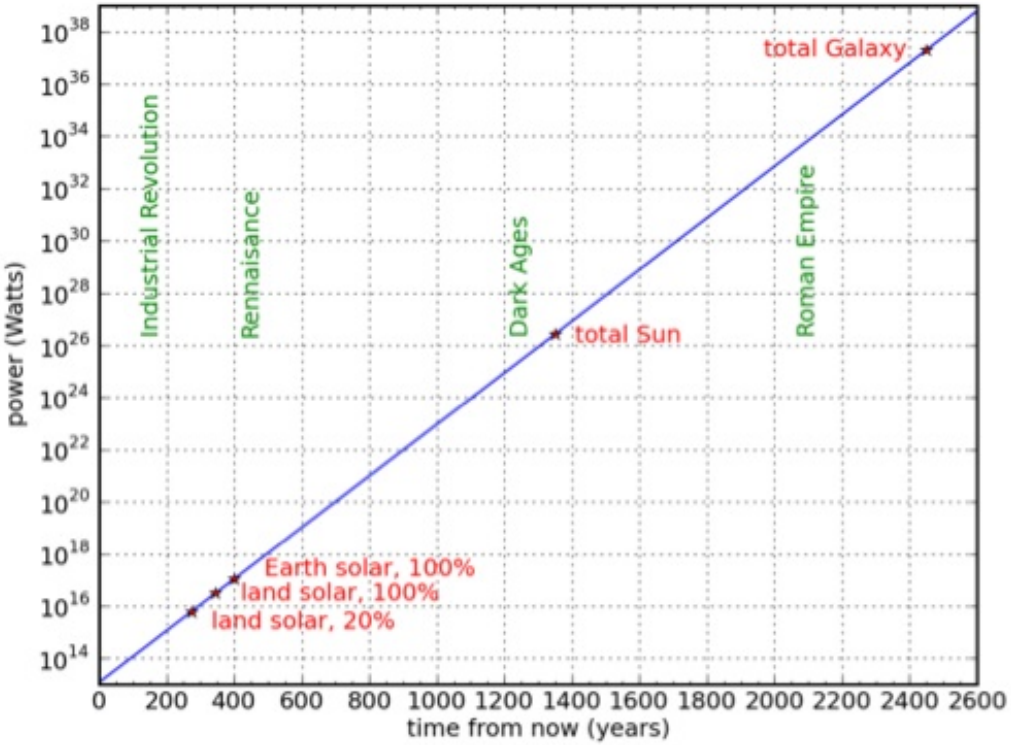
- What is the growth rate?
- By how much does it increase every 100 years?
- What energy sources could satisfy the future demand?

More Questions

- Current Demand = 12 TW

- Sources:
  - Sun (when and how will we outgrow it?)
  - Nuclear: fission + fusion (can we?)
  - Tidal + Geothermal (Negligible)

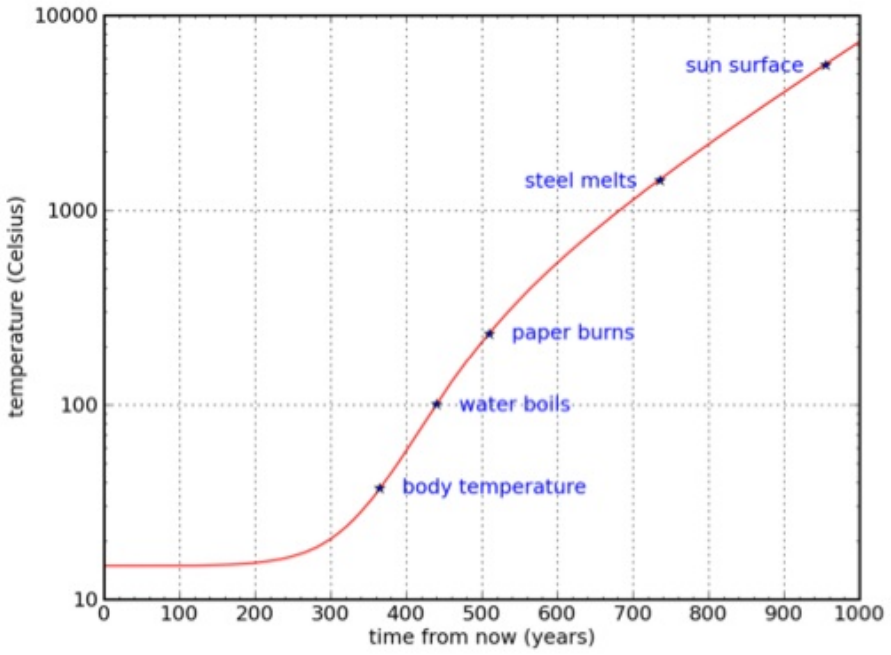
Using Solar Energy



— Global power demand under sustained 2.3% growth on a logarithmic plot. In 275, 345, and 400 years, we demand all the sunlight hitting land and then the earth as a whole, assuming 20%, 100%, and 100% conversion efficiencies, respectively. In 1350 years, we use as much power as the sun generates. In 2450 years, we use as much as all hundred-billion stars in the Milky Way galaxy.

Source: Tom Murphy

Temperature Rise for Earth



— Earth surface temperature given steady 2.3% energy growth, assuming some source other than sunlight is employed to provide our energy needs and that its use transpires on the surface of the planet. Even a dream source like fusion makes for unbearable conditions in a few hundred years if growth continues. Note that the vertical scale is logarithmic.

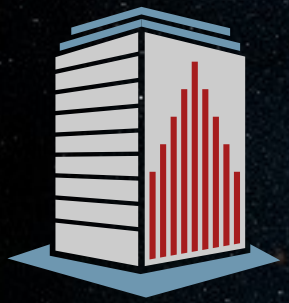
Source: Tom Murphy

The End

QUESTIONS?

@bergesmario

marioberges.com



INFERLab

Intelligent Infrastructure Research Laboratory

Carnegie Mellon