**Proceedings of the ASME 2018 Power and Energy Conferece**
**PowerEnergy2018**
**June 24-28, 2018, Lake Buena Vista, FL, USA**

**PowerEnergy2018-7327**

# DEMOGRAPHICAL ENERGY USAGE ANALYSIS OF RESIDENTIAL BUILDINGS

**Alice Sokolova**
Electrical and Computer Engineering Department
San Diego State University
San Diego, California 92182
Email: asokolova@sdsu.edu

**Baris Aksanli**
Electrical and Computer Engineering Department
San Diego State University
San Diego, California 92182
Email: baksanli@sdsu.edu

## ABSTRACT

Residential energy consumption constitutes a significant portion of the overall energy consumption. There are significant amount of studies that target to reduce this consumption, and these studies mainly create mathematical models to represent and regenerate the energy consumption of individual houses. Most of these models assume that the residential energy consumption can be classified and then predicted based on the household size. As a result, most of the previous studies suggest that household size can be treated as an independent variable which can be used to predict energy consumption. In this work, we test this hypothesis on a large residential energy consumption dataset that also includes demographic information. Our results show that other variables like income, geographic location, house type, and personal preferences strongly impact energy consumption and decrease the importance of household size because the household size can explain only 26.55% of the electricity consumption variation across the houses.

**Keywords:** residential energy consumption, demographic analysis, data clustering, modeling

## INTRODUCTION

Residential energy consumption constitutes a large portion, around 40%, of the overall U.S. energy consumption [1]. Therefore, it is important to analyze this big consumption and understand what the main factors affecting it. There are several studies that collect residential data, including energy consumption, household demongraphics, etc. Some examples include 1) American Time Use Survey [2], that includes demographic information regarding residential households across the U.S. but no energy consumption data (works using similar datasets from France [3], U.K. [4] and Spain [5], 2) Residential Energy Consumption Survey [6] include residential energy consumption information but the data granularity is very coarse and demographic information is limited, 3) MIT REDD [7] and Smart* [8] datasets include detailed energy consumption data but no demographics information for a very limited set of houses. As can be seen, the previous datasets do not provide a good opportunity to combine energy and demographic analysis in a fine-grained manner. As a result of this, most of the time, studies have assumed that there is a correlation between household size and energy consumption, but the nature of this correlation has not been investigated.

Many studies identify a correlation between household size (and/or other demographics factors, [9], [10], [11], [12], [13]) and energy consumption. For example, a big electricity utility company, SDG&E (San Diego Gas & Electric), provided data shown in Figure 1 [14]. The data confirms the correlation between energy use and household size. However, in order to determine whether the correlation is also a dependency, further analysis must be performed. A correlation can imply in misleading conclusions due to multicollinearity. Examples include the likelihood of higher income correlating with more appliances, and flats tending to be smaller than detached houses, introducing a confounding between dwelling type and size" [13]. Previous studies, [13] from England, [9] from the US, [10] from France, carry out a similar study to ours but either the appliance and power meter data they use relies heavily on surveys, or they do not use fine-grained time-series power consumption data to capture patterns. In contrast, we use fine-grained appliance energy

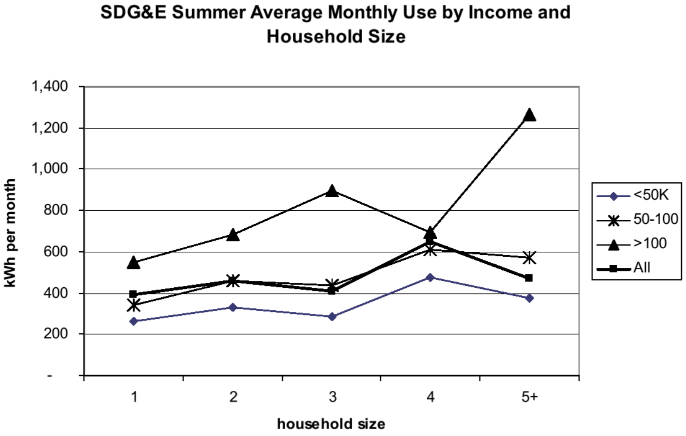


**FIGURE 1**. SDG&E summer average monthly use by income and household size [14]

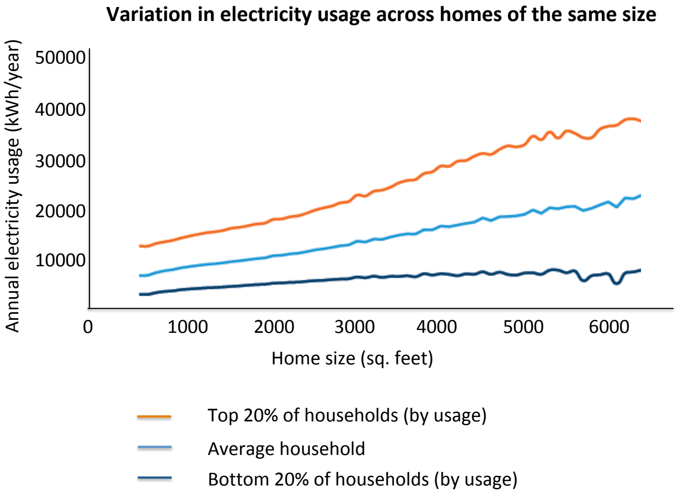


**FIGURE 2**. Variation of electricity usage across homes of the same size. Image recreated from [16].

consumption data that is reported every 1 minute or 15 minutes intervals. Also, in order to draw accurate conclusions about the influence of household size on energy consumption, other variables must be controlled for to the highest extent possible. This study attempts to isolate household size from other predictors.

In order to carry out our study, we leverage data provided by Pecan Street Inc. [15]. Pecan Street provides high resolution power consumption data, every 1-minute and 15 minutes, along with demographic information regarding the participating residential houses. Pecan Street collects data from participating homes, which participate in surveys and agree to have elecctricity egauges installed. As a result, Pecan Street provides access to residential electricity consumption data paired up against the socio-demographic data of the residents. The level of detail found in the survey is substantial. The survey includes crucially important questions such as resident number by age, total annual income, education level, retrofits, heating and cooling systems, total number of certain appliances, and more. The biggest advantage of this dataset is that the appliance usage does not depend on surveys but fine-grained data, which helps us create a very detailed clustering mechanism that includes up to 40 dimensions.

Of the nearly 1000 homes listed in Pecan Street's database, we analyze over 200 dwellings. The homes chosen for analysis are the ones which have complete or mostly complete electricity data and survey data for the year 2014. All selected homes are either single-family houses or town houses. Most of them are located in Austin, TX. However, others are located all over the U.S., such as San Diego CA, Dallas TX, and Richmond WA.

We choose the house type as the control variable. One can suggest that house size is one of the more important energy consumption predictors. Data collected by Opower, a residential energy management service, confirms that energy consumption increases for larger sized homes [16], shown in Figure 2.

We make another assumption that the annual income variable and the house size variable are closely related in a way that they are interchangeable. This is because house size can be seen as a consequence of income. Furthermore, other consequences of income, such as appliance type and number have less influence on energy consumption than house size. Because of these reasons, we choose the home size, rather than annual income, as the main control variable. Research performed at Arizona State University confirms that house size and total annual income are positively correlated [17], which is also outlined in Figure 3.

Our experimental results, testing the correlation between energy consumption and household size, show that other variables like income, geographic location, house type, and personal preferences strongly impact energy consumption and decrease the importance of household size. Thus, in the discussion section, we suggest that household size is not an independent variable and is not an accurate predictor of a home's energy consumption. We also present more detailed approach to modeling a household's energy consumption using other variables and factors, such as time-dependency, appliance-specific consumption, etc.

In summary, our paper makes the following contributions:

1. We develop a systematic method to correlate residential energy consumption to demographic information. We use classification methods that use up to 40-dimensional data to better capture the dependencies among different variables.
2. Our study uses high-granularity, time-series data to represent energy consumption at both house level and individual appliance level. This helps us to capture time-based energy usage patterns.
3. We demonstrate that survey based studies might overestimate their explanatory power because surveys might not represent accurate energy usage behavior and patterns.

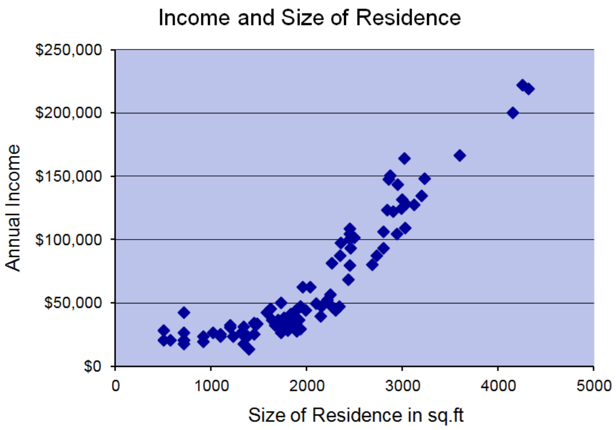


**FIGURE 3**. Income vs. size of residence [17]

## METHODOLOGY

Our goal in this study is to predict the energy usage of a household if its demographics are known. But to accomplish this, we take the converse approach. A statistical analysis is performed on the energy use data available from Pecan Street database in order to develop an algorithm capable of correctly identifying family type (number of household members) based on energy usage.Our starting point is to first observe and analyze the total monthly electricity use per household. By far the two most common statistics for data representation are mean and standard deviation. We separate hourly electricity use data, available from the egauge meters, into months and extract the mean and standard deviation. We follow this procedure in order to capture the inherent, time-dependent characteristics of the energy consumption time-series data.

In order to separate the resulting data into categories (from here on called ”clusters”), we use the k-means clustering algorithm [18]. K-means clustering is a comparative iterative algorithm which divides input data into groups based on similarity. During each iteration of the algorithm, each data point is reassigned to the nearest centroid and the new centroid is calculated for the new cluster formed during the iteration. The k-means clustering algorithm has no definite dimensionality limit, which makes it especially useful for the purpose of our study. Initially, we input the two-dimensional data every household has two statistics associated with it: monthly average and standard deviation, to the algorithm. Figure 4 shows the result of the clustering (based on average and std. dev.), leading to 4 groups, represented by different colors.

Average use and standard deviation follow an approximately linear relationship: more use means higher standard deviation of use. If energy use were indeed a function of household size, this graph would make intuitive sense. The actual data set indicates that largest size group is the ”couples” group. Households inhabited by couples are about three times as numerous and households with one or three members, while households with more than three members become less numerous. Specifically, in the data set 30 households are single-resident, 94 are couples, 34 households have three members, 34 households number four people, and only 16 houses are home to families of five or more.

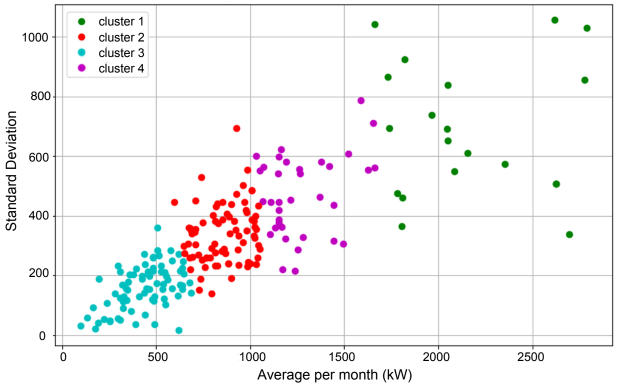


**FIGURE 4**. Clustering results based on only mean and standard deviation of monthly energy consumption

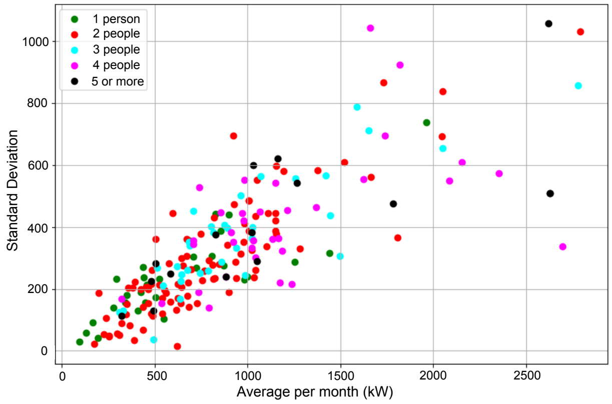


**FIGURE 5**. Clustering results based on only household size, energy consumption is monthly

By visually inspecting Figure 4 and assuming a correlation between use and household size, it is reasonable to hypothesize that couple-households correspond to the blue and red points on the graph. The lower left blue data points might indicate single households. Purple and some red data points likely show households with four people, and the 19 green data points seem to correlate fairly well with the 16 households numbering five or more people. However, Figure 5 displays the actual distribution of average and standard deviation corresponding to actual household size.

It is evident that there is indeed no similarity between the clusters generated and the actual distribution of use according to household size. It seems as if the mean and standard deviation

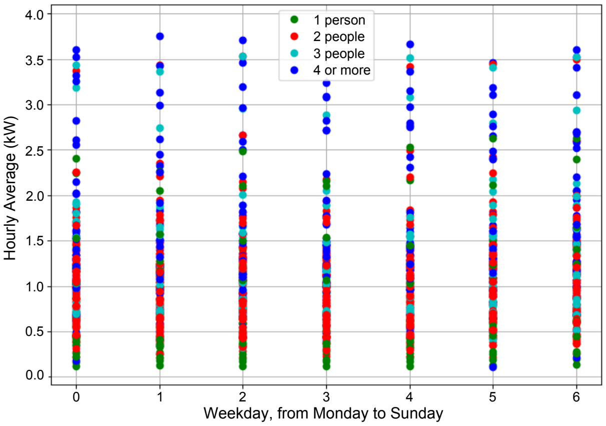


**FIGURE 6**. Clustering results based on daily average energy consumption for different family sizes

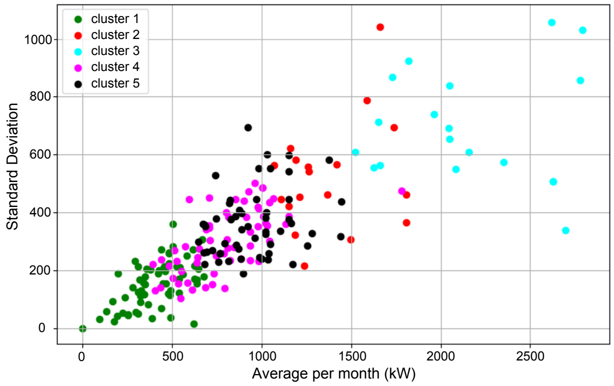


**FIGURE 7**. Clustering results with 20-dimensional data, energy consumption is monthly

of each house is completely random. However, it is worth noting that, although heavily scattered, there is still some positive dependence on household size. The lowest mean monthly use belongs to a house with a single resident. The highest standard deviation belongs to a house with five or more members. Red data points, indicating couples, tend to gravitate toward the center, while blue and purple data points dominate over red ones toward the left side of the plot. It is also important to note that black data points, indicating households with 5 or more, can be found across the whole range of energy use, from about 300 W to 2500 W. Clearly, we need to take a more advanced approach in order to attempt to replicate the real distribution pattern.

Our next attempt is to analyze behavioral patterns based on days of the week. For every household, we plot the average energy use for each of the seven days in the week and analyze them. We immediately abandoned this approach as it seemed unpromising (therefore, we do not include the results of that analysis). In Figure 5, a correlation can be visually observed, even though it is heavily scattered. Whereas the results of separating data into days of the week showed no such correlation. There are few logical reason to hypothesize that different household sizes will display clear behavioral patterns during the seven days of the week. Analysis showed that, indeed, we should not pursue this approach.

Similarly, analyzing average energy use during different times of the day proved equally unpromising. In our results, outlined in Figure 6, it can be noted that red data points (families of 2 people) tend to gravitate toward lower averages, and blue data points (families of 4 or more people) gravitate toward higher averages, but there is no noticeable dependence on the weekday. Thus, we have chosen to continue applying to approach of k-means clustering. Pecan Street data makes it possible to look at energy consumption of individual household appliances. It is reasonable to hypothesize that different family sizes use appliances differently. To implement this analysis, we choose nine appliances, for which the most data was available, for analysis. These appliances are: air conditioner, clothes washer, dishwasher, clothes dryer, furnace, microwave, lights, garage, and refrigerator. The total use was still included in the analysis, which resulted in ten data arrays to be used as input. The monthly average and standard deviation was calculated for each appliance, giving a total of twenty data points for every household. Thus, twenty-dimensional data was fed into the k-means clustering algorithm. Results are pictured in Figure 7.

These results are much more promising because the scattering effect noted in Figure 5 is starting to emerge in Figure 7. Thus, we use the same approach in the further expansions to the clustering algorithm. Encouraged by the improved accuracy, we implement two major modifications in the next version. Firstly, we perform the analysis in greater detail by pulling hourly data, instead of monthly summations. This was done in order to note behavioral patterns, which may have been obscured by looking at monthly cumulative values. Secondly, we increase the dimension of the data again by adding another statistic in addition to average and standard deviation: entropy (information theoretic).

Entropy is a quantity related to the probability distribution of an event. The value of entropy, H, increases as the uncertainty of the value increases. H is zero if an event is completely predictable. Applied to this study, entropy is a useful statistic to use because the data we use is by nature random and unpredictable. Entropy provides a numerical quantity to describe this unpredictability. Thus, the entropy statistic measures how ”random” is the energy use of a single household. The hypothesis is that the ”randomness” of energy use will vary for different family sizes. For example, a large family may have a more chaotic and unpredictable schedule. However, a single resident might be more likely to have a consistent schedule of energy use. This is also the reasoning behind using hourly energy data rather than monthly data.

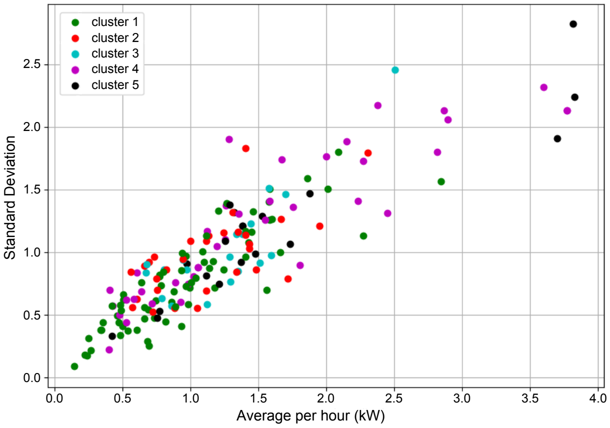


**FIGURE 8**. Clustering results with 40-dimensional data, energy consumption is hourly

With the previously mentioned changes implemented, the data fed into the k-means clustering algorithm is now forty-dimensional. For every household, we generate an input array which contains the average, standard deviation, time of day value, and entropy for all ten appliances used for analysis. As before, total summed energy use is included in the list of appliances. Results are shown in Figure 8.

Of all statistical analysis attempted, this variation of the algorithm provides the best results. We performed other experiments to improve this algorithm by changing the statistical operations performed on the data. The options we tried include changing the ”entropy” function to ”skew”, ”harmonic mean”, ”k-statistic”, and more. We have also tried to replace the time category with another statistical measure. We observe the best results by using the previously described algorithm, and we provide more details about its results in the following section.

## RESULTS

Figure 8 shows the plot of the best clustering generated by the k-means algorithm. A visual inspection confirms a similar scattering pattern to the actual distribution. Even more promising are the cluster sizes formed. The dataset we use indicates that there are 30, 92, 34, 34, and 15 households corresponding to the 5 family size groups, ranging from one resident to five plus residents. The cluster sizes generated by the algorithm are 33, 99, 36, 19, and 18. These numbers correlate well with the family size groups. Next, we are going to determine whether the real family size groups overlap with the generated clusters. Table 1 presents a comparison between generated clusters and actual data.

Unfortunately, the clusters do not overlap as much as desired. In fact, it is impossible to definitively map the generated clusters onto the real data. Highlighted in yellow and blue cells indicate maximum overlap. Ideally, the yellow and blue highlights should have occupied the same cells, and there should have been a total of five highlighted cells. If a correspondence between real size groups and generated clusters could be identified, even if it was partially erroneous, the experiment would have been successful. However, an overlap occurred only between the first cluster and the two-resident family group, probably by nature of their much larger size in comparison to other clusters. At best, it is possible to identify the most accurate correspondence possible between the real and generated clusters. The best fitting correspondence is identified in Table 2. Highlighted in yellow is the permutation of clusters and residents which yields the highest percent overlap. At the highest, the success rate (cluster overlap between generated clusters vs. actual households) is 26.55%.

| Residents | Clusters | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | Total |
| 1 | 13 | 6 | 2 | 7 | 2 | 30 |
| 2 | 44 | 13 | 10 | 18 | 7 | 92 |
| 3 | 18 | 8 | 2 | 2 | 4 | 34 |
| 4 | 17 | 5 | 4 | 3 | 5 | 34 |
| 5+ | 7 | 1 | 1 | 6 | 0 | 15 |
| Total | 99 | 33 | 19 | 36 | 18 | 205 |

**TABLE 1**. Comparison between generated clusters and actual household sizes

| Residents | Clusters | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 43.33% | 20% | 6.67% | 23.33% | 6.67% |
| 2 | 47.83% | 14.13% | 10.87% | 19.57% | 7.61% |
| 3 | 52.94% | 23.53% | 5.88% | 5.88% | 11.76% |
| 4 | 50% | 14.71% | 11.76% | 8.82% | 14.71% |
| 5+ | 46.67% | 6.67% | 6.67% | 40% | 0% |
| Average percentage overlap | | | | | 26.55% |

**TABLE 2**. Overlap percentage: residents vs. clusters

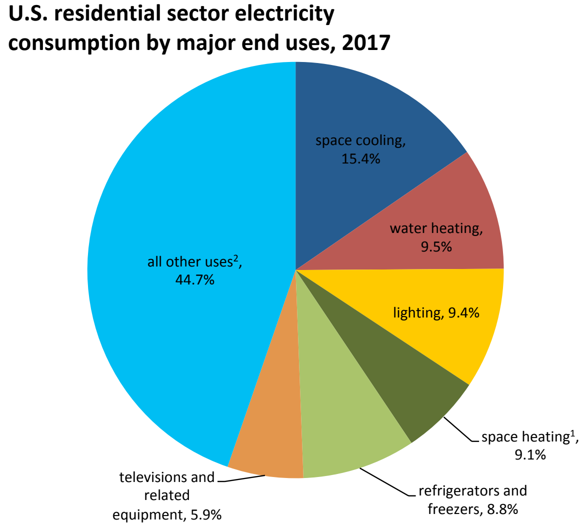
## DISCUSSION

This study proves that total energy consumption of a household somewhat depends on family size, but not predictably. Too many other factors, equally unpredictable, prevent simple modeling of a household’s energy consumption. This section will discuss possible reasons why household size explains only a quarter of energy use variation. Additionally, we compare our study with other similar studies in Table 3, in terms of their predictors, data sources, and the obtained explanatory powers.

In a previous study, Huebner et al. from UCL [13] found that Socio-demographics variables on their own explained about 21% of the variability in electricity consumption with household size the most important predictor.. Their research yielded similar results as this study (in comparison to our 26.5% best overlap

| Study | Country | Predictors used | Data source | Explanatory power (%) |
|---|---|---|---|---|
| Huebner et al. [13] | England | Household size, resident income and age | Survey | 21 |
| Bartiaux and Gram-Hanssen [21] | Denmark | Household size (detached houses) | Survey | 27.6 |
| | Belgium | Household size (detached houses) | Survey | 4.8 |
| Nielsen [22] | Denmark | Household size, appliances, and floor area | Survey | 64 |
| Genjo et al. [23] | Japan | Appliances | Survey | 60 |
| Our work | USA | Household size, time of day, appliances | Time-series, digital data | 26.55 |

**TABLE 3**. Comparison between our work and related studies



Notes:
[1] Includes consumption for heat and operating furance fans and boiler pumps.
[2] Includes miscellaneous appliances, clothes washers and dryers, computers and related equipment, stoves, dishwashers, heating elements, and motors not included in the uses listed above.
Source: U.S. Energy Information Administration, *Annual Energy Outlook 2018*, Table 4, February 2018

**FIGURE 9**. U.S. residential sector electricity consumption by major end uses in 2017. Image created using the table in [19].

ratio). Another study by Marcus and Ruszovan found that Increases in family size are associated with only relatively modest impacts on usage below $100,000 in income and in the cool climate zone. [14]. The UCL Energy Institute study claims that appliance use explains a much larger share of energy use variation than household size does [13]. In this section, we examine the relationship between household size and appliance use.

The U.S. Energy Information Administration determined that residential energy use is described by the pie chart in Figure 9 [19]. Arguably, some of the appliance categories influencing electricity consumption have a limited dependence on household size. For example, according to the U.S Energy Information Administration 2009 Residential Energy Consumption Survey [6], most households own only one refrigerator, which is plugged in continuously every day all year [20]. The energy consumption of food refrigeration, covering 8.8% of total residential household energy use, really does not depend on family size. Space heating and cooling, totaling 26.6% of household energy use, also has a limited dependence on family size because a heating and cooling system maintains a set temperature in the whole house regardless of how many people inhabit the home. It can be argued that house size depends on resident number, and heating in turn depends on house size, but the relationship is not direct.

Examining separate household appliances shows that the dependence on appliance use on household size is not logically obvious. Many factors must be considered, and crucially important among them is the unpredictable factor of person preferences. Different families use their appliances very differently, adding a whole level of random variables. In order to achieve accuracy in predicting energy use, the model of residential electricity consumption must be very complex and include many input variables, not only household size.

Furthermore, our study also demonstrates the importance of using time-series, digital data, rather than relying on survey data. This is because survey data might not completely represent the human behavior that fully affects the energy consumption patterns in a house. For example in Table 3, some previous studies obtained high explanatory power using survey data, which is not the case when we made a similar analysis with the time-series, digital data demonstrating real usages.

## CONCLUSION

This study analyzes the relationship between residential energy consumption and household sizes. We demonstrate that a successful model that can demonstrate the relationship between the energy consumption and household demographics should include a detailed analysis that incorporates more than household size as the sole analysis focus. Our study shows that even a 40-dimensional energy consumption clustering method that incorporates detailed individual appliance usage analysis can explain 26.55% of the variation in energy consumption across the houses. Therefore, the analysis must be in more detailed. In particular, three branches of analysis must be considered: socio-demographical data, house structure data, and geographic data. Within the socio-demographic branch.Although previous studies obtain similar results, analyzing appliance ownership and use, socio-demographic variables, building variables and selfreported energy-related behavior, their appliance usage data includes mainly survey-based methods and does not include detailed, fine-grained appliance usage information. In our work, we include numerical data from hundreds of houses and apply rigorous modeling and data clustering methods to understand the relationship between residential energy consumption and household demographic information. The main result of our study is that residential energy consumption will always have a significant random component that cannot be directly modeled by demographical information, but rather daily human activities.

## REFERENCES

[1] United States Energy Information Administration, 2017. How much energy is consumed in US residential and commercial buildings? https://www.eia.gov/tools/faqs/faq.php?id=86&t=1. Accessed: December 2017.

[2] Bureau of Labor Statistics, 2014. American Time Use Survey. https://www.bls.gov/tus/. Accessed: December 2017.

[3] Basu, K., Hawarah, L., Arghira, N., Joumaa, H., and Ploix, S., 2013. "A prediction system for home appliance usage". *Energy and Buildings,* ***67***, pp. 668–679.

[4] Collin, A., Tsagarakis, G., Kiprakis, A., and McLaughlin, S., 2012. "Multi-scale electrical load modelling for demand-side management". In IEEE PES ISGT Europe.

[5] López-Rodríguez, M., Santiago, I., Trillo-Montero, D., Torriti, J., and Moreno-Munoz, A., 2013. "Analysis and modeling of active occupancy of the residential sector in spain: an indicator of residential electricity consumption". *Energy Policy*.

[6] United States Energy Information Administration, 2009. Residential Energy Consumption Survey. https://www.eia.gov/consumption/residential/. Accessed: December 2017.

[7] Kolter, J. Z., and Johnson, M. J., 2011. "Redd: A public data set for energy disaggregation research". In Workshop on Data Mining Applications in Sustainability (SIGKDD), San Diego, CA, Vol. 25, pp. 59–62.

[8] Barker, S., Mishra, A., Irwin, D., Cecchet, E., Shenoy, P., and Albrecht, J. "Smart*: An open data set and tools for enabling research in sustainable homes". *SustKDD'12*.

[9] Elnakat, A., Gomez, J. D., and Booth, N., 2016. "A zip code study of socioeconomic, demographic, and household gendered influence on the residential energy sector". *Energy Reports,* ***2***, pp. 21–27.

[10] Hache, E., Leboullenger, D., and Mignon, V., 2017. "Beyond average energy consumption in the French residential housing market: A household classification approach". *Energy Policy,* ***107***, pp. 82–95.

[11] Zhang, M., Song, Y., Li, P., and Li, H., 2016. "Study on affecting factors of residential energy consumption in urban and rural Jiangsu". *Renewable and Sustainable Energy Reviews,* ***53***, pp. 330–337.

[12] Lévy, J.-P., and Belaïd, F., 2017. "The determinants of domestic energy consumption in France: Energy modes, habitat, households and life cycles". *Renewable and Sustainable Energy Reviews*.

[13] Huebner, G., Shipworth, D., Hamilton, I., Chalabi, Z., and Oreszczyn, T., 2016. "Understanding electricity consumption: A comparative contribution of building factors, socio-demographics, appliances, behaviours and attitudes". *Applied Energy,* ***177***, pp. 692–702.

[14] Marcus, W. B., and Ruszovan, G., 2007. "Know your customers: A review of load research data and economic, demographic, and appliance saturation characteristics of california utility residential customers". *prepared on behalf of The Utility Reform Network for California Public Utilities Commission App*, pp. 06–03.

[15] Pecan Street Incorporation, 2015. Dataport. http://www.pecanstreet.org/category/dataport/. Accessed: December 2017.

[16] OVO Energy, 2014. Whats the average gas bill and average electricity bill in the UK? https://cleantechnica.com/2013/03/08/us-electricity-consumption-much-more-evenly-distributed-than-income-wealth/. Accessed: December 2017.

[17] Waissi, G. Multiple linear regression CASE. http://www.public.asu.edu. Accessed: December 2017.

[18] Hartigan, J. A., and Wong, M. A., 1979. "Algorithm as 136: A k-means clustering algorithm". *Journal of the Royal Statistical Society. Series C (Applied Statistics),* ***28***(1), pp. 100–108.

[19] US Energy Information Administration, 2018. Annual Energy Outlook 2018 - Table 4: Residential sector key indicators and consumption. https://www.eia.gov/tools/faqs/faq.php?id=96&t=3. Accessed: February 2018.

[20] United States Energy Information Administration, 2012. Annual energy review - household end uses: fuel types, appliances, and electronics, selected years, 1978-2009.

[21] Bartiaux, F., and Gram-Hanssen, K., 2005. "Socio-political factors influencing household electricity consumption: A comparison between denmark and belgium". In ECEEE Summer Study Proceedings, Vol. 3, pp. 1313–25.

[22] Nielsen, L., 1993. "How to get the birds in the bush into your hand: results from a danish research project on electricity savings". *Energy policy,* ***21***(11), pp. 1133–1144.

[23] Genjo, K., Tanabe, S.-i., Matsumoto, S.-i., Hasegawa, K.-i., and Yoshino, H., 2005. "Relationship between possession of electric appliances and electricity for lighting and others in japanese households". *Energy and Buildings,* ***37***(3), pp. 259–272.