International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Special Issue NCTFRD18

Machine Learning Techniques for Load Disaggregation in Rural off-Grid, Isolated Solar Systems

[1]Thara D.K, [2]Priya M, [3]Indira K.M

[1]Assistant Professor, [2,3]UG Students, Department of ISE, CIT, Gubbi Tumkur, India.

manjunathpriya477@gmail.com

Abstract— **In power systems and electricity markets, accurate monitoring of electricity demand is important in order to manage real time load balancing and for distribution and transmission planning. In the context of rural electrification, the uncertainty of both supply and demand is large. And the central reason that the solar-based micro grids and home systems can be expensive. Oversizing battery storage and solar panel size or limiting the electricity available to users can reduce this uncertainty but increases system costs. In ideal scenario data should be captured on a per-household level and must be measured on a sufficiently fine time resolution to translate the demand into individual, user activities. This massive quantity of house hold demand data requires automatic tools to convert time series power usage data into data on the use of individual appliances. This paper presents methodology, based on data acquired in individual, isolated solar home systems in Jharkhand, India. On utilizing classification and clustering algorithms to create activity-based models that can be used to conduct load forecasts. Additional statistical data analysis can yield insights on users' power consumption behavior in relation to exogenous variables such as time of day and conditioned on ambient air temperature.**

Keywords – Machine Learning, classification, solar home system.

I. INTRODUCTION

The lack of electricity is one of the most pressing concerns in developing world. This deficiency impedes most aspects of human development; health, education and economic development. In developing countries, grid electricity is often unreliable or unavailable. The governments of these countries do not have the financial resources to increase generation to meet increasing demand, let alone to electrify off-grid areas. Also, grid extension to small remote areas can be very expensive.

Individual systems (such as solar home systems and diesel generators) have seen growth in recent years due to ease of deployment. However, they are very expensive and require complex financing solutions. Moreover, it is difficult to extend their operation. Recently, micro-grids have received more attention, especially in the developing world, due to their relatively low cost of electricity achieved by aggregate generation. Solar-based micro-grids have been recognized as a key enabler of electricity provision to the over one billion people living without energy access to-date. Despite significant cost reductions in solar panels, micro-grids can still be expensive for number of reasons. Yet, the recent advent of intelligent devices low-cost computation can enable lower cost of micro-grid architecture through demand-side management and network control. In any power system, the design and costs is directly related to expected demand to be served in an area. In context of developing world, having accurate estimation of electricity consumption is of critical importance in order to inform both electrification planning efforts and real-time balancing of supply and demand. This is consistent and recurring challenge faced by practitioners when providing electricity access in rural, off-grid areas. Broadly speaking, there are two main methods –qualitative and quantitative – for estimating demand profiles when designing power systems for rural electricity access. In former method, one such approach is based on country level or regional level demographics – using indicators such as household income, population density, poverty levels, etc. to estimate and predict household electricity demand per year. Another qualitative approach based on individual survey questionnaires to obtain a baseline understanding of income levels, power levels, of appliances desired, typical daily usage patterns, etc. Once such information is obtained by individuals this data is obtained can be used to create average daily load profiles- which can be extrapolated through basic randomization methods over entire year. In the developing world, high resolution (i.e. household level at sub-hourly time intervals) data sets are rarely available but can extremely useful. Such data sets can also validate conventional demand modeling approaches during design faces of electricity planning efforts. Over- or under-estimation of demand has an effect on both cost and reliability of solar-based micro-grids or solar home systems. The intension of this effort is to understand the process of acquiring data, applying activity-based clustering

 DOI : 10.18231/2454-9150.2018.0809

techniques and generating statistical models for household level electricity consumption.

II. Data Procurement and Processing

Data sets on electricity consumption in developing world, especially in rural, off-grid areas, are not widely available. So related works conducted in [3] and [4] where driving motivation in acquiring raw electricity consumption data sets. In coordination with Tata Steels Rural Development Society[TSRDS], data loggers were installed in protected enclosures and connected to solar home systems(powering lights and fans) in three rural houses without electricity access outside the city of Jamshedpur, located in India state of Jharkhand.

Over six months of current, voltage, and temperature data are acquired with temporal resolution of 1 minute. These data sets are used in analysis conducted and presented herein; Three data loggers were installed in solar home systems in particular village. Each household wherein a data logger was installed were constructed of dried mud with metallic rooftops and was without windows, this allowed the interiors of the house to be slightly cooler than the ambient temperature but also made homes quite dark. The solar home systems installed in these rural households were designed and made by an off-grid solar home systems provider in India. The system installed in each household included with a 70W (12V) solar panel, a 75Ah (12V) battery, four 9W CFL lights and one 14W fan – all operated in DC and connected to a charge controller that was hung and from the wall in each household. Data was successfully time stamped and stored at a 1-minute sampling rate and uploaded, yielding over 6 months electricity consumption data: examples of household's consumption data at different time frames are shown in Fig. 1:

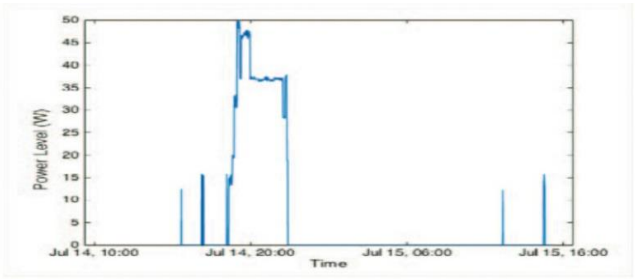

Fig. 1(A): Example of Daily Load Profile

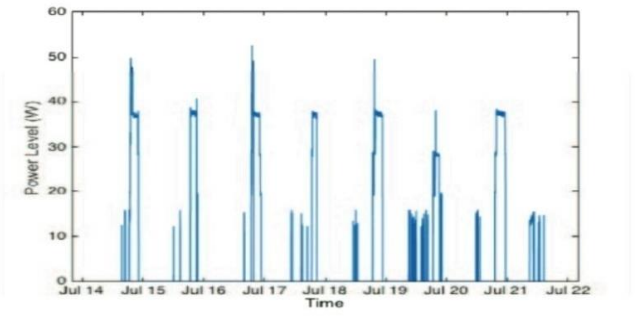

Fig. 1(B): Example of Weekly Load Profile

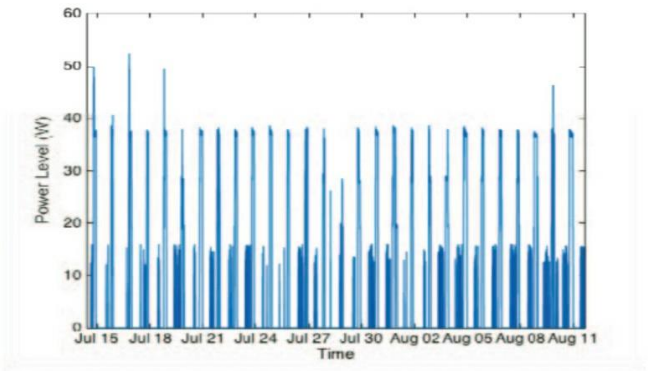

Fig. 1(C): Example of Monthly Load Profile

As depicted in Fig.2, Each data logger used for this research purpose was on off-the-shelf 16-bit voltage data logger. In order to measure current, a shunt resistor was used in series with each households charge controller output – and converting from the 65,536 available digital reading values gave a resolution of ~.31mV. Current and voltage were recorded for both solar panel and battery, along an on – board thermistor to measure ambient temperature.

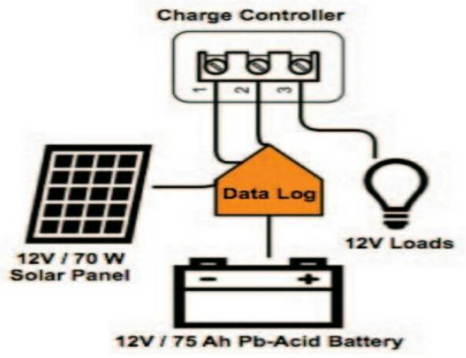

Fig. 2(A): Layout of Data Logger Connection

From the raw data sets, the amount of energy consumed (in the unit watt – minutes) could be determined by first multiplying the respective instantaneous current and voltage measurements from the solar panel and the battery. For any given data point, the amount of solar power plus battery must equate to the amount of load consumed or energy spilled ('spillage'). Thus, to separate the load from the spillage the following heuristic was used: if the inferred solar irradiance was <10W\m^2, which could roughly be determined as the nameplate panel capacity, efficiency, and thus area were all known, then the value was likely an evening load. To separate the expected daytime load in the fan from the amount of energy spilled if the value was within the confidence interval of the fans expected power consumption's of 14W, the value was classified as load.

III. Load Disaggregation Approach

When considering disaggregation approaches based on load data, key parameter such as AC versus DC, sampling rate, data resolution (i.e. appliance, household or feeder) must be taken into account. In the situation at hand, the number – and thus potential combination – of appliances is known and network operates in DC; what is not known at this point

 DOI : 10.18231/2454-9150.2018.0809

is the standard deviation of each appliance. The goal is to identify user's state (i.e. combination of appliances that is on) given a power consumption reading: In order to access performance of clustering algorithms, labeled truth must be generated for this data set at minute level. Given the information in Table 1, which is organized in increasing magnitude of power consumed / number of appliances on, labeled truth is generated by using method in taking the means of the adjacent states as upper and lower bounds: having this range allows for values to be classified based on power reading into its equivalent state at each minute. This process is done with original data set with one minute resolution, as conducting this process after conducting hourly average would loss not only granularity, but also distort power levels to be unexpected values or perceived as alternate states.

Table 1: State & Appliance Combination

State	Expected Mean Power (W)	Appliance Combination
1	0	None
2	9	1 CFL
3	14	1 Fan
4	18	2 CFL's
5	23	1 CFL + 1 Fan
6	27	3 CFL's
7	32	2 CFL's + 1 Fan
8	36	4 CFL's
9	41	3 CFL's + 1 Fan
10	50	4 CFL's + 1 Fan

IV. Clustering Algorithm Performance And Result

In machine learning, clustering algorithms are used to identify grouping and assigning labels to data points one challenge with clustering algorithm is the need to know the number of possible clusters (or in this case, states) in advance of applying the clustering algorithms. As shown above, the number of possible states with data sets at hand is 10 for one particular household. However, plotting the frequency of states occurred based on generate4d truth data, it's important to note that users may not reach all the possible states – some without any sort of frequency or recurrence. Specifically, three or four states are rarely, if ever, reached for any household (such as states 7, 9, 10 as defined above in the case of this particular household and shown in Fig 3.). The reasons for these maybe unable to supply the requisite discharge rate of current to reach these higher power states.

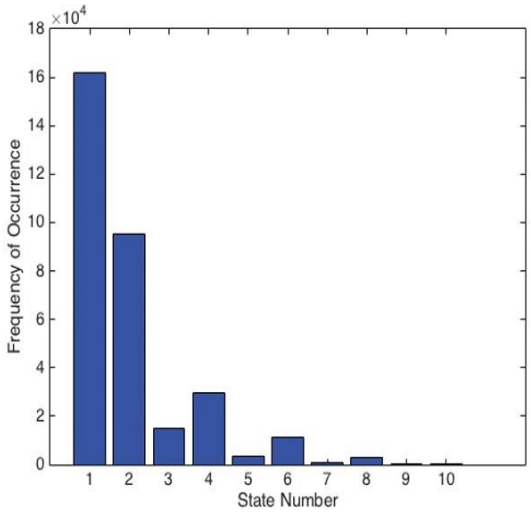

Fig. 3: Frequency of Occurred States (Minutes)

Both K–Means and Gaussian Mixture Model (GMM) are applied to these data sets at the household level. The K-Means algorithm is used to identify clusters embedded in data sets and assign data points to the particular clusters, but is non-probabilistic in its nature. The algorithms objectives are to minimize the total sum of the square of distances between each data point and its assigned cluster. GMM is related but different in that the clusters are probabilistic and normally distributed around cluster centers; thus, each data point is assigned a probability that it belongs to a given cluster. Both these algorithms do not have Closed–formed solutions, and rely on iterative Expectation – Maximization (EM) algorithm to recursively determine the clusters centers (and associated mixing coefficient in the case of GMM) until convergence criterion is reached. As the results below indicate, the algorithms are used in number of different scenarios to access their respective performances in assigning user states (i.e. combination of appliances on) based on reading in power consumption level. First, Fig.4 shows how K-Means algorithm performance when used with 7 or 10 clusters, and with or without initialization of centroids. For k=10, initialization does help the algorithm converge towards centroid values that are expected for each of the states – and to a lesser extent for k=7. Comparing between k=10 and k=7, table 2shows that the algorithm had the latest aggregate sum of distances – and number of iterations required for convergence – from points to centroids when initialized for k=10. This result was initially unexpected, such as states with expected higher magnitude centroids such as 9 and 10, are rarely ever reached. However, given that the majority of the states are reached in lower magnitude centroids, such as states 1 through 4, having more clusters available allow for the algorithm to space out centroids – which makes them closer to the expected centers when compared to k=7.

DOI : 10.18231/2454-9150.2018.0809

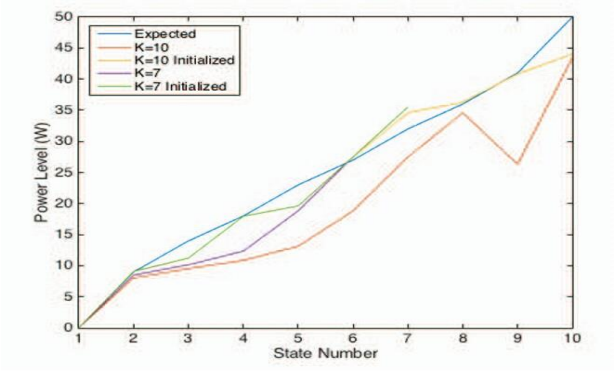

Fig. 4: K-Means Centroid Results

Table 2: K-Means Algorithm Performance

	Num. Iterations	Total Sum of Distances
K=7	32	8.8×10^4
K=7 (w/ initialization)	23	1.2×10^5
K=10	66	4.2×10^4
K=10 (w/ initialization)	23	1.1×10^5

Second, Fig.5 indicates how Gaussian Mixture Model performs with both k=7 and k=10; the key difference with K-Means is that here, the algorithm assigns a probability that a given point belongs to a cluster i.e. normally distributed around its center. Appendices B and C includes table showing the difference in the values of the expected centers and mixing proportions when the GMM is or is not initialized with expected mean values. It's clear that initialization as a substantial impact on the algorithms performance – as the GMM centers and mixing proportions are very similar to the expected values. Without initialization, the majority of the GMM centers hover around zero, with the sum of mixing properties of the first five centers equating 0.513, which expected mixing proportion for one centroid expected to be at zero. For k=10, a nearly identical story holds, illustrating the importance of initialization and trade of maximizing the distance between centroids – while maximizing total distance between data points and centroids.

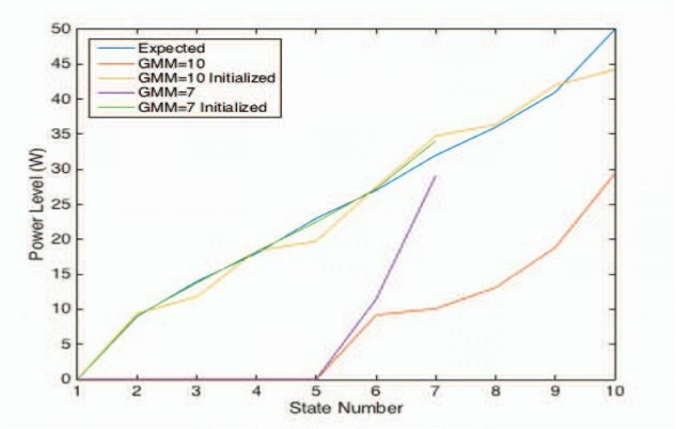

Fig. 5: GMM Center Results

As the figures above indicate, the performance of the clustering algorithm in terms of the centers' proximity to expected mean power levels of states is far better when initializing with the expected levels of each state; without this, the algorithm tends to congregate near zero, given that is where over 50% of the power reading lie (for this particular household, through similar results hold for the other two households). This is a rational enough conclusion with a labeled data sets and knowledge of appliance states beforehand.

Note that the above analysis is based on the algorithms' performance over the entire data set of length 319,680 data points (222 days x 24 hours x 60 minutes). For situation in the field where neither the expected power consumption of the appliances or the number of appliances is not known in advance, different algorithms may be required. But another point to understand is how the algorithms fare after seeing subsets of the data - after just a day, a week, or a month. Tables 3 and 4 below indicates the performances of GMM for K=7 and K-Means for K=10, both with initialization of centers, after seeing data sets for length 1,440 (24 hours x 60 minutes), 10,080 (24 hours x 60 minutes x 7 days), and 40,320 (24 hours x 60 minutes x7 days x 4 weeks) values:

Table 3: Data Set Length Effect on K-Means

	One Day	One Week	One Month
No. of Iterations	11	14	13
Centers	0	0	0
	9.309	9.12	9.19
	13.72	13.25	12.90
	19.05	19.12	19.25
	28.18	20.57	22.34
	37.33	27.99	28.33
	49.04	28.64	31.17
	50.69	37.28	37.44
	52.18	47.97	48.24
	53.98	53.84	53.80

Table 4: Data Set Length Effect on GMM

One Day	Mixing Proportion	One Week	Mixing Proportion	One Month	Mixing Proportion
0	0.613	0	0.711	0	0.706
9.31	0.183	9.14	0.098	9.21	0.99
13.72	0.012	13.34	0.016	12.93	0.043
19.05	0.043	19.13	0.034	19.26	0.019
19.06	0.000	21.23	0.000	22.52	0.001
28.18	0.033	28.18	0.027	28.33	0.025
28.19	0.000	30.32	0.000	31.38	0.001
37.33	0.074	37.28	0.084	37.44	0.093
37.39	0.000	47.04	0.005	48.25	0.009
52.43	0.042	53.83	0.024	53.82	0.006

The above analysis yields insight on a number of timely and important topics. First, a process for data acquisition of residential load profiles is shared herein, which can be replicated by practitioners aspiring to get a better understanding of electricity consumption patterns in rural off-grid contexts. Second, the analysis applies and compares a pair of clustering algorithms commonly used in fields of machine learning. A key learning is a effect of initializing the algorithms with the expected clusters centers – as the combination of the noise embedded in a raw data and that appliances are not on during majority of the time adversely effects the algorithms' performance. Further more understanding of these techniques' convergence based on the amount of data they receive, along with the effect of number of

 DOI : 10.18231/2454-9150.2018.0809

clusters are also shared. Third, assuming the capability for load disaggregation in place, this paper show the potential for deeper inside into the users' electricity consumption patterns. Though this data is only available at three households to-date, this type of analysis is used to challenge and corroborate assumption on end user's electricity consumption

V. Conclusion And Future Work

The above analysis yields insight on a number of timely and important topics. First, a process for data acquisition of residential load profiles is shared herein, which can be replicated by practitioners aspiring to get a better understanding of electricity consumption patterns in rural off-grid contexts. Second, the analysis applies and compares a pair of clustering algorithms commonly used in fields of machine learning. A key learning is a effect of initializing the algorithms with the expected clusters centers – as the combination of the noise embedded in a raw data and that appliances are not on during majority of the time adversely effects the algorithms' performance. Further more understanding of these techniques' convergence based on the amount of data they receive, along with the effect of number of clusters are also shared. Third, assuming the capability for load disaggregation in place, this paper show the potential for deeper inside into the users' electricity consumption patterns. Though this data is only available at three households to-date, this type of analysis is used to challenge and corroborate assumption on end user's electricity consumption behavior in these contexts. Future work should extend to remote monitoring and rely on low-cost, open-sourced hardware – connecting metering devices to a cloud-based infrastructure for real-time analytics; but given day-to-day variability across individual household, tailored probabilistic, state-based approach that can take into account exogenous variables – such as input/output. As intelligent devices, network, and micro-grids begin to penetrate in off-grid areas, the intent to gather quantitative data for monitoring electricity consumption must be accompanied with useful algorithm and visualization techniques. To drive insight. However, as has been stated before, access to data sets of user's electricity consumption is not readily available in this space. Generating and sharing data sets will give further information about electricity planning efforts in regards to generation asset, sizing, performance real-time load disaggregation algorithms.

References

[1] Varun Mehra, Rajeev Ram, Claudio Vergara " A Novel Application of Machine Learning Techniques for Activity-Based Load Disaggregation in Rural Off-Grid, Isolated Solar Systems".

[2] W. Inam et. al. "Architecture and System Analysis of Micro-grids with Peer-to-Peer Electricity Sharing to Create a Marketplace which Enables Energy Access".

[3] D. Soto and V. Modi. "Simulations of Efficiency Improvements using Measured Microgrids Data".

[4] M. Buevich et. al. "Fine-Grained Remote Monitoring, control and Pre-Paid Electrical Service in Rural Micro-grids".

[5] C. Bishop. "Pattern Recognition and Machine Learning".

VI. Appendices

Appendix A: K-Means Centroid Results

State	Expected Mean	K=10 Centroid	K=10 Centroid (w/ initialization)	K=7 Centroid	K=7 Centroid (w/ initialization)
1	0	0	0	0	0
2	9	8.06	9.09	8.51	9.08
3	14	9.54	11.23	10.15	11.23
4	18	10.87	17.96	12.36	17.96
5	23	13.14	19.62	18.81	19.62
6	27	18.81	27.49	27.49	27.49
7	32	27.49	34.59	35.52	35.52
8	36	34.60	36.29		
9	41	36.32	40.85		
10	50	43.45	44.05		

Appendix B: GMM Results, K=7

State	Expected Mean	Original Mixing Proportions	GMM Centers (w/ initialization)	GMM Mixing Proportions	GMM Centers (w/o initialization)	GMM Mixing Proportions
1	0	0.5107	0	0.4882	0.0323	0.1113
2	9	0.3528	9.09	0.3167	0.0323	0.1113
3	14	0.0422	13.79	0.0726	0.0323	0.1113
4	18	0.0727	18.27	0.0725	0.0323	0.0849
5	23	0.0027	22.47	0.0204	0.0323	0.0942
6	27	0.0342	27.23	0.0021	11.44	0.4393
7	32	0.0018	34.05	0.0075	29.08	0.0479
8	36	0.0097				
9	41	0.0003				
10	50	0.0000				

Appendix C: GMM Results, K=10

State	Expected Mean	Original Mixing Proportions	GMM Centers (w/ initialization)	GMM Mixing Proportions	GMM Centers (w/o initialization)	GMM Mixing Proportions
1	0	0.5107	0	0.5107	0	0.1021
2	9	0.3528	9.39	0.2835	0	0.1021
3	14	0.0422	11.83	0.0844	0	0.1021
4	18	0.0727	18.35	0.0506	0	0.1021
5	23	0.0027	19.74	0.0248	0	0.1021
6	27	0.0342	27.49	0.0349	9.18	0.1391
7	32	0.0018	34.75	0.0069	10.10	0.2045
8	36	0.0097	36.38	0.0039	13.11	0.0244
9	41	0.0003	41.98	0.0001	18.81	0.0753
10	50	0.0000	44.26	0.0002	29.42	0.0460

 DOI : 10.18231/2454-9150.2018.0809