uSIM2020 - Building to Buildings: Urban and Community Energy Modelling, November 12th, 2020

System of statistical approaches for community energy demand modelling

Dr Sandhya Patidar[1*], Dr David P. Jenkins[2], Dr Andrew Peacock[2], Dr Kumar Biswajit Debnath[2], Dr Peter McCallum[2]
[1]Institute for Infrastructure & Environment,
[2]Institute for Sustainable Building Design,
School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh, UK
(*Corresponding Author: S.Patidar@hw.ac.uk)

Abstract

Modelling transient community-level peak energy demand event is often challenging, as it requires the acquisition and systematic analysis/modelling of electricity demand data across a large number of buildings. Electricity demand data with diverse demand characteristic can be analysed/modelled/aggregated (in time) to understand the impact of various micro-level activities (specifically, peak demand household-level activities occurring simultaneously across multiple dwelling at a specific time) on the community-level demand curve. However, in real-life applications, the availability of good-quality electricity demand data across a large number of multiple dwellings within a community is often challenging.

This paper is aimed to investigate the potentials of k-means clustering approach for developing a systematic sampling, weighting and demand aggregation strategy for projecting community-level demands with high precision, just by using a small sample of buildings and easily accessible contextual information (e.g. average monthly demand or various activity periods during a day). These selected samples of dwellings are processed with a novel system of demand synthesis model developed by authors, referred to as HMM_GP. Five different variants of k-means clustering are developed using statistical mean, median and proportion of demand during four different periods of days. Corresponding to each variant five aggregation schemes are constructed. The HMM_GP model is underpinned by a *hidden Markov model (HMM)* for simulating synthetic demand and a *Generalised Pareto (GP) distribution* to effectively model dynamics of peak demand events. Aggregation schematics are demonstrated for 30-minutely demand dataset collected over four weeks in July 2017 for 74 dwellings for a case-study community of Fintry (Scotland).

Introduction

Conventionally, monthly/quarterly meter readings were utilised by professionals and researchers to construct empirical electricity demand curves (Suganthi and Samuel 2012, Bhattacharyya and Timilsina 2009, Bhattacharyya and Timilsina 2010). These curves are widely used for understanding various temporal demand characteristics, however, with limited scope to accommodate any quality assurance for a highly uncertain future. This level of information is now in-adequate for addressing various challenges, uncertainty and complex issues (e.g. climate change, technological changes, policies, economy, infrastructure development,

behavioural changes, etc.), that are at the heart of any long-term energy-related planning and sustainability discussions of a rapidly evolving modern society (McKenna, et al. 2018). With growing interest in energy-focused community-level projects, not just at the local levels but at the global level as well, a large volume of electricity demand data (e.g. from the installation of smart meters) is now available. Smart meters are generating a large volume of data and can be used to extract various useful information/patterns for supporting various decision-making and planning activities to benefit both industry and society (Stankovic, et al. 2016). This can be achieved by improving the capabilities of existing approaches and, also by developing new efficient approaches for processing/analysing these large volumes of information/data in manageable/effective way. There is also a desire for the UK to lead innovation within these sectors (Industrial Strategy: building a Britain fit for the future 2017).

In this context, this paper is aimed to demonstrate the potentials of a widely applied k-means based clustering approach (J. A. Hartigan 1975, Hartigan and Wong 1979) in developing efficient demand aggregation strategies. The underpinning idea is *to select a considerably small sample of size 15% of individual demand profiles (informed by the k-means clustering using simple monthly-level statistics as characterising features) that can be processed with highly-efficient synthetic demand simulation models (such as the HMM_GP* (Patidar, et al. 2019) *which can be applied to high-resolution individual demand profiles), to generate aggregated demand profiles with high accuracies*.

Need for such a model arises from various perspective, for example, in practical applications, smart meter data are still not available for several communities. In such a case, data collected for a small sample of buildings along with some monthly/quarterly-level statistics can be used to generate high-resolution aggregated demand profiles. Even for the communities/regions where smart meters are rolled out, simultaneous availability of ***good-quality*** electricity demand data, across a large number of buildings within the community (required for developing aggregated demand profiles) is often practically challenging. This is not just an issue for

developing/under-developed nations but is also a potential challenge for many developed nations. Nevertheless, the proposed approach can be also used to plan strategies for future roll-out of smart meters in developing/under-developed nations to selected buildings and thus optimising the resource allocation (Kshetri and Voas 2018).

The next section will give an outline of the dataset and case-study used in the paper. Rest of the paper is organised to give an overview of research methodology, analysis, key results, and discussion.

Case study

To develop and demonstrate the proposed methodology, a case-study community, Fintry (Smith 2018, Howell 2020) is selected. Fintry is a beautiful village in Stirlingshire, Central Scotland, surrounded by the Endric water, Fintry hills and the Campsie Fells. Fintry embraces approximately 350 households (c700 inhabitants) and is an off-gas grid community that mainly uses electric, oil and LPG based heating.

Data Organisation

With an inspiration to transform Fintry a carbon-neutral and sustainable community, Fintry Development Trust has been set up in 2007, which now has almost 250 active members. As part of various projects, FDT has contributed to the installation of various forms of renewable energy generation plants (e.g. Solar PV, Biomass and Wind) in the community. In particular, as part of the SMART Fintry project, funded from the Scottish Government Local Energy Challenge Fund (LECF), electricity demand data for 115 dwellings in the community at a temporal resolution of 30-minutes for almost a year were collected (Smith 2018). Following a thorough pre-analysis, a portfolio of 74 dwellings for July has been identified as good-quality (continuous dataset with less than 5% of missing values) dataset for the present study. Missing values are infilled using a logical algorithm developed by the authors and detailed as a step-by-step procedure elsewhere (Debnath, et al. 2020).

Research Methodology

Research methodology involves the integration of K-means clustering with the HMM_GP model. K-means approach is used for identifying a small suitable sample for constructing aggregation schematics and the HMM_GP model is used for generating synthetic demand from the sample. The HMM_GP model is highly technical involving a system of statistical approaches for processing high-resolution demand data. Further, details on the underpinning methodology for the HMM_GP model can be referred elsewhere (Patidar, et al. 2019). This paper will mainly focus on developing and identifying a suitable variant for K-means clustering.

K-means Clustering

A clustering approach is aimed to organise/partition the large collection of cases/items into a disjoint groups/clusters such that the items belonging to a cluster are alike to each other for some specified features/characteristics than the items in other clusters. Key underpinning methodology for constructing aggregation schematics here is based on a k-means clustering approach. K-means clustering performs unsupervised learning task to organise the collection of items across a fixed number (k) of clusters by optimising the squared error function (Nisbet, Miner and Yale 2018). Key steps of the k-means clustering algorithm are briefly discussed as below and can be referred elsewhere (Larose and Larose 2014):

a) Identify an optimum number of clusters. Elbow method is applied

b) Allocate k items (randomly) as initial cluster centres (centroid/mean).

c) Allocate each item to their nearest cluster centre depending on "nearest" distance criterion (e.g. Euclidean distance [1]), thus creating k clusters, $C_1, C_2, \dots, C_k$.

d) Identify cluster centre (centroid), which is the mean value of all data point in the cluster [2] and update the cluster centre.

e) Repeat above steps iteratively to minimise *mean squared error* (MSE) and until the algorithm converges, i.e. when cluster centre does not change or when no significant reduction is observed in MSE.

Five variants of K-means clustering are constructed at five different periods: i) T_1- 00:00 to 23:30; ii) T_2- 00:00 to 06:00; iii) T_3- 06:30 to 12:00; iv) T_4- 12:30 to 16:30; v) T_5 - 17:00 to 23:30. Six statistics used for K-means clustering: i) mean (m_i), ii) median (m_d) [i-ii) both estimated at each of the five distinct periods T_i for $i \in 1, 2, \dots, 5$); iii) proportion (p_2) of demand during the period T_2 to the total demand(p_1) during the entire day T_1; iv) proportion (p_3) of demand during the period T_3 to p_1; v) proportion (p_4) of demand during the period T_4 to p_1; vi) proportion (p_5) of demand during the period T_5 to p_1. Mean (m_i), median (m_d), and P_i for $i = 1, 2, \dots, 5$, are measured for 30-minutely demand date for each of the 74 dwellings and for the entire duration of July.

Simulation and Analysis

K-means clustering is applied using the functions "fviz_cluster", available in R package "*factoextra*". Further details on the theory and application of k-means procedure including its implementation in R can be referred elsewhere (Kassambara 2017).

[1] Other distance measurements are Squared Euclidean, Manhattan distance, Pearson correlation distance, Spearman correlation distance, Kendall correlation distance, Chebyshev.

[2] For n data points $(a_1, b_1, \dots m_1), (a_2, b_2, \dots m_2), \dots (a_n, b_n, \dots, m_n)$, centroid is found as $\left(\frac{\sum a_i}{n}, \frac{\sum b_i}{n}, \dots, \frac{\sum m_i}{n}\right)$.

The clustering variant I, represents the simplest possible grouping of 74 dwelling that includes basic and widely applied statistics, mean and median demand measured over the period T_1 for the entire duration of July.

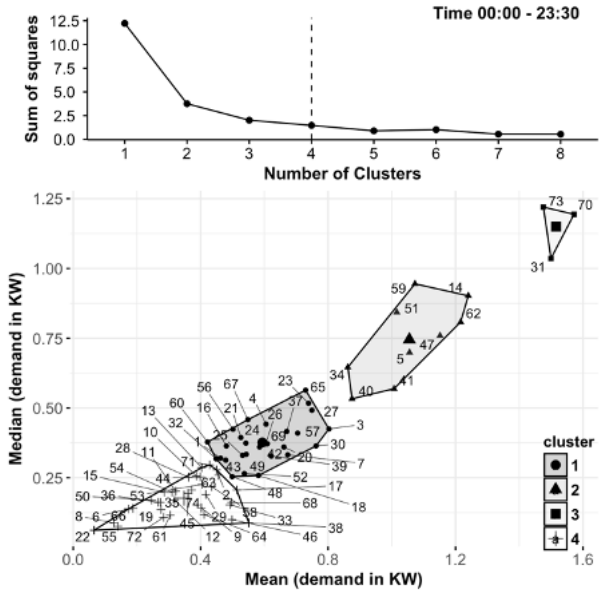

Figure 1: Elbow plot and k-means clustering variant I.

Figure 1 illustrates the cluster analysis for the variant I. Top panel presents the plots of Elbow method, applied for obtaining an optimum number of cluster. The Elbow method estimate and plots the total within-cluster sum of squared errors (WCSS, explained later) for the different number of clusters. An optimum number of clusters is chosen when a change in WCSS is not significant for change in the number of clusters. Elbow method suggested 4 clusters as the optimum choice for clustering variant I.

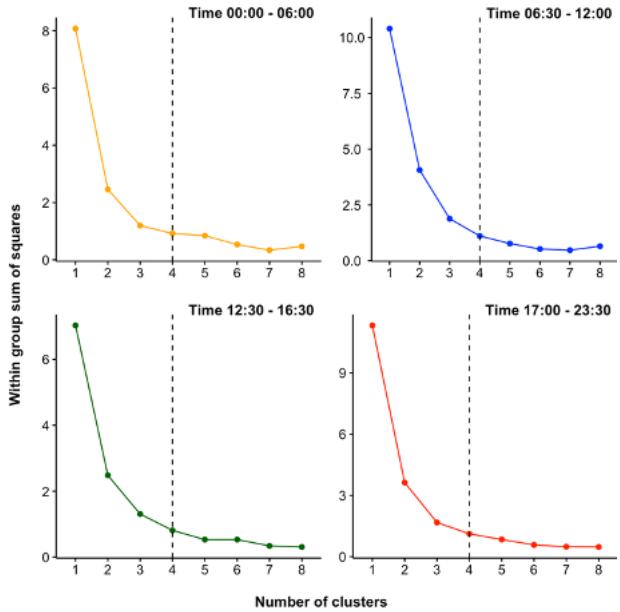

Figure 2: Elbow plot for k-means clustering variant II, III, IV and V for the time period T_2, T_3, T_4 and T_5, respectively labelled in the plot.

'Bottom panel' of Figure 1, shows the distribution of 74 dwellings across the four clusters in variant 1. All the four clusters are not overlapping and show a considerable variation between them. Dwellings in cluster 3 have considerably high mean and median than the dwelling in cluster 4. Further statistics and performance indicators for cluster variant 1 can be examined in Table 1 and 2.

The architecture of the other four variants is mostly similar in the sense that they all are performing grouping based on mean and proportion of demand at four distinct periods of the day. Selection of time period is intended to reflect a different level of activities occurring across the day, specifically to investigate the hypothesis "*k-means clustering if involves the feature that accounting in time periods with a comparatively large volume of peak demand activities can more effectively predict peaks demand in aggregated profiles*". From a pre-analysis of observed aggregated demand profiles, it appears that time periods T_2 and T_4 has comparatively low peak demand activities, whereas most of the peak demands (activities) are observed in the time period T_3 and T_5.

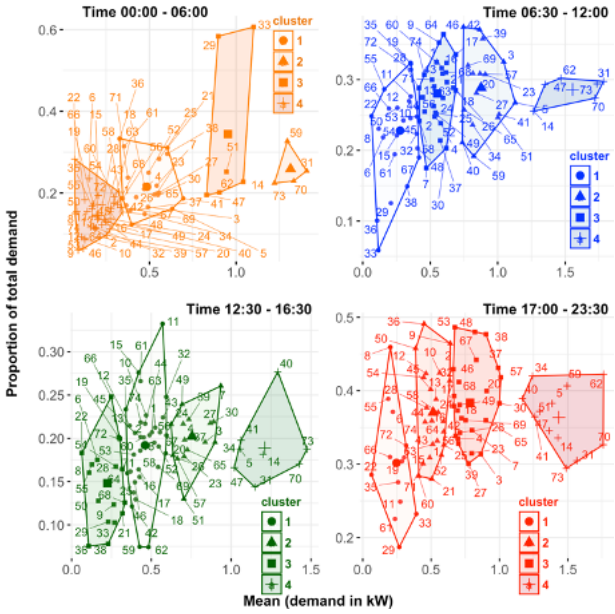

Figure 3: K-means clustering of variant II, III, IV and V.

Figure 2 illustrates the Elbow plot for clustering variant II, III, IV and V for the other four distinct time periods period T_2, T_3, T_4 and T_5, respectively. Interestingly, for all the four variants, Elbow plot suggested an optimum cluster size of 4. Figure 3 shows the simulation results of k-means clustering for all the four variants II, III, IV and V. From the visual inspection of Figure 4, all the four variants were successful in organising 74 dwellings in four distinct, non-overlapping, compact clusters. For all the cluster variants (II-V), x-axis scales comparatively in the ranges of a mean demand of 0.1-1.7 kW whereas y-axis appears to vary in the range of 0.1 to 0.4 for variant III and IV and in the range of 0.0 – 0.6 for variant II and V.

Table 1 presents three key overall performance indicators for the five variants of k-means clustering described

above. These are i) *Total sum of square* (TSS) measures the total variance in the data; ii) Total *withing-cluster sum of square* (WCSS) measures total withinness of clusters, as the average squared distance of all point within a cluster; and iii) *Between-cluster Sum of Squares* (BCSS) measure the average squared distance between all the centroids.

Table 1: Performance indicator for k-means clustering

K-means variant	**I**	**II**	**III**	**IV**	**V**
Statistic used for clustering	m_1, m_d	m_2, p_2	m_3, p_3	m_4, p_4	m_5, p_5
TSS	12.22	8.07	10.40	7.03	11.32
WCSS	1.37	0.86	1.10	0.81	1.12
BCSS	10.85	7.20	9.30	6.22	10.20

On the performance scale, a low score of TSS is an indicator of the less total variance in the overall process. Table 1 indicates that the cluster variant IV attains the minimum variance, which is closely followed by variant II. Notably, both the variants II and IV are designed using the periods with comparatively less peak demand activities. Further on assessing WCSS, the smallest score that indicating less variance within the cluster (i.e. compact clustering) is achieved for cluster variant IV which is again closely followed by cluster II. Thus, a low score for WCSS indicates both the variants IV and II are comparatively performing better than variant III, V and I.

Table 2: Performance indicator for clusters

Cluster	**1**	**2**	**3**	**4**
K-means variant I for the time period $\boldsymbol{T_1}$ 00:00 – 23:00				
Count	28	9	3	34
Total demand	12364	7065	3383	8793
Mean	0.59	1.05	1.52	0.35
Median	0.38	0.74	1.15	0.18
WCSS	0.44	0.31	0.02	0.6
K-means variant II for the time period T_2 00:00 – 06:00				
Count	23	4	6	41
Total demand	11121	4181	4115	12187
Mean	0.49	1.31	0.95	0.20
Avg. Proportion	0.22	0.26	0.34	0.15
WCSS	0.30	0.03	0.24	0.30
K-means variant III for the time period T_3 06:30 – 12:30				
Count	24	16	28	6
Total demand	5270	9365	10453	6067
Mean	0.28	0.88	0.55	1.55
Avg. Proportion	0.23	0.29	0.28	0.29
WCSS	0.39	0.25	0.22	0.25
K-means variant IV for the time period T_4 12:30 – 16:30				
Count	34	15	17	8
Total demand	12350	7946	3961	7348
Mean	0.46	0.75	0.22	1.21
Avg. Proportion	0.19	0.20	0.15	0.19
WCSS	0.28	0.16	0.15	0.22
K-means variant V for the time period T_5 17:00 – 23:30				
Count	13	24	25	12
Total demand	2491	7400	11267	10448
Mean	0.25	0.51	0.78	1.43
Avg. Proportion	0.30	0.37	0.38	0.36
WCSS	0.14	0.19	0.37	0.42

*Total demand is measured in kWh for all 74 dwellings in entire July 2017.

Finally, on accessing BCSS, smallest score is achieved for variant IV and II. A high score of BCSS indicates a good separation between the different clusters, thus according to BCSS based assessment, cluster I is performing best. Though it should be noted that total variance is also highest for clustering variant I, so these values are expected to be consequently high for variant I. To further assess and compare these five clustering variants, performance indicators specific to individual clusters for all the five variants are detailed in Table 2. Interestingly, as expected, the average proportion of demand is highest for cluster variant V for all the four clusters. Information presented in Table 2 is used to design aggregation schematics (discussed in the next subsection).

Designing Aggregation schematics

The cluster-specific measurements, such as counts, total demand, and average proportion are used to draw sample dwellings from the clusters and for designing a systematic logical aggregation schematic. A sample size of 15% of 74 (i.e. total number of dwellings) ~ 11 dwellings is chosen. Five aggregation schematics corresponding to five cluster variants are designed and thoroughly analysed. For each of the five distinct clustering variants, a selection of 11 sample dwellings is conducted from the four different clusters using a logical cluster weighting formula, given as below:

$$W_c = s * average(P_S, P_D, P_A),$$

where, W_c is the number of dwellings to be selected from a cluster c, s is the total size of the sample required, P_s is the proportion of cluster size to the total number of dwellings (i.e. 74), P_D is the proportion of total demand accounted in the cluster c, P_A is the average proportion of the cluster. For the purpose of demonstration, the sample selection procedure for Aggregation schematic 2 (corresponding to cluster variant II) is presented here. In Aggregation schematic 2, the number of dwellings to be sampled from cluster 1, 2, 3 and 4 are respectively estimated as:

$$W_1 = 11 * average\left(\frac{23}{74}, \frac{11121}{31605}, 0.22\right) = 3.23 \sim 3;$$

$$W_2 = 11 * average\left(\frac{4}{74}, \frac{4181}{31605}, 0.26\right) = 1.63 \sim 2;$$

$$W_3 = 11 * average\left(\frac{6}{74}, \frac{4115}{31605}, 0.34\right) = 2.02 \sim 2;$$

$$W_4 = 11 * average\left(\frac{41}{74}, \frac{12187}{31605}, 0.15\right) = 3.99 \sim 4.$$

Further, Figure 3 has been used to randomly select sample dwellings from each of the clusters to ensure selection is considerably varied. Please note that, for aggregation schematic 1, an only average of P_s and P_D are used. The

same procedure is applied to select a sample of 11 dwellings for each of the five aggregation schematic.

Synthetic simulation of demand

For all the five different Aggregation schematics, aggregated demand profiles are compiled by generating a suitable number of synthetic demand profiles through the application of the HMM_GP model. The sample of dwellings selected from each of the clusters (as specified above) is processed with the HMM_GP model to generate the required number of synthetic demand profiles. With reference to the above example, in aggregation schematic 2, a sample of three dwellings selected from cluster 1 is simulated using HMM_GP model to generate 23 synthetic demand profiles. 30 minutely observed demand profiles over the entire July are used for all the three samples and 8 synthetic demand profiles (30 minutely over the entire July) are generated corresponding to each of the observed sample dwellings. With the same analogy, a total of 4, 6 and 41 synthetic demand profiles are generated using HMM_GP model from the sample of 2, 2 and 4 observed sample profiles drawn from the cluster 2, 3 and 4 respectively.

Result Analysis (Aggregation Schematics)

To assess the potential impacts of different clustering variants in designing an efficient aggregation schematic (with capabilities in effectively estimating peak demands), a thorough performance analysis of all the five Aggregation schematics is performed. Figure 4 presents a visual comparison of aggregated demand profiles (composed by temporal addition of 74 observed 30-minutely individual demand profiles) with the 74 synthetically constructed aggregated demand profiles for a week. Synthetic aggregated demand profiles are constructed by adding individual synthetic demand profiles generated from HMM_GP model for the sample of 11 dwellings (samples selected using clustering-based information as specified above).

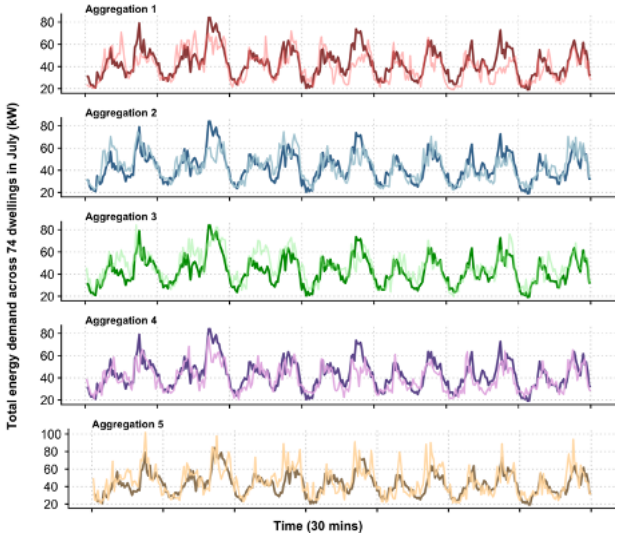

Figure 4: Comparing observed dynamically aggregated demand profiles for 74 dwellings for a week in July (dark thick lines) with synthetically generated aggregated demand profiles (lightly shaded line) for five Aggregation schematics.

In Figure 4, observed aggregated demand profiles are presented using thick dark lines ('*Maroon*' for Aggregation 1, '*Navy*' for Aggregation 2, '*Green*' for Aggregation 3, '*Violet*' for Aggregation 4, and '*Brown*' for Aggregation 5). Corresponding synthetic aggregation profiles are presented with the light shade of thin lines ('*Pink*' for Aggregation 1, '*blue*' for Aggregation 2, '*light green'* for Aggregation 3, '*purple*' for Aggregation 4, and '*yellow*' for Aggregation 5). Visually all the aggregation scheme appears to perform reasonably good in capturing transient dynamics of observed aggregated profiles. However, on closely observing their behaviour for capturing the dynamics of peak demands it appears that Aggregation scheme 5 often over-estimate large peaks.

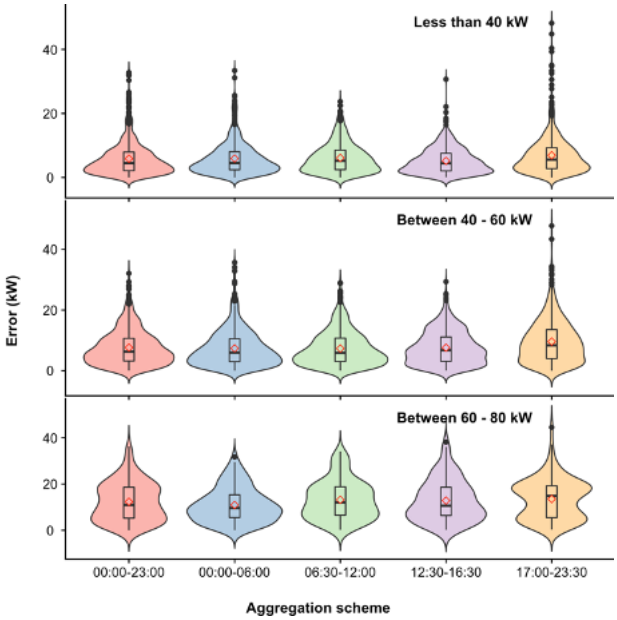

Figure 5: Comparing violin plots for assessing error distribution for five aggregation schematics in predicting observed aggregated demand in three ranges.

To further assess the model performance a thorough statistical analysis of aggregation error term is conducted. Aggregation error at time instance t is defined as $E(t) = O(t) - S(t)$, where $O(t)$ is the observed aggregated demand at time t and $S(t)$ is the synthetic aggregated demand at time t. Errors are estimated for each of the half-hourly aggregated demand values for the July and error analysis is performed for three observed aggregated demand value (ranges): a) Less than 40 kW, b) between 40-60 kW and c) between 60-80 kW. Results are illustrated in Figure 5 using the violin plots.

Violin plots are comprehensive illustrations used to present the density distribution of data along with the box plot (Hintze and Nelson 1998). A box plot presents the five summary statistics, minimum, 25th percentile (1st quartile), 50th percentile (Median), 75th percentile (3rd quartile) and Maximum. In addition to the box plot, a violin plot also provides a kernel probability density (frequency) distribution of data at different values on the

y-axis. The violin plots presented here are generated in R-package 'ggplot2' (Pedersen 2020). In each of the violin plots, box plots are marked as boundaries (in dark black lines) outlining a box shape. Lower and Upper bars on the box presents (1st and 3rd quartiles respectively) interquartile ranges, a dark line passing within the box indicate the median value, and a red empty star symbol indicating mean of the error distribution. Extended lines from the box emanating from lower and upper quartile indicates the variability of data and reach to mark the minimum and maximum values respectively. Black solid filled dots are outliers. Area enveloping the box plot (forming a violin-like structure) and filled in with different colours here shows the kernel density distribution of error. A '*pink*' infill is used for demonstrating error distribution for Aggregation 1. Similarly, a '*blue*' infill is used for Aggregation 2, '*light green*' for Aggregation 3, '*light purple*' for Aggregation 4, and '*yellow*' is used for Aggregation 5.

Figure 5 'Upper panel' shows the violin plot of error distribution corresponding to aggregated load values in the range of less than 40 kW. 'Middle panel' shows results for load values in the range of 40-60 kW and 'Bottom panel' shows results for load values between 60-80 kW.

For less than 40 kW (Figure 5, Upper Panel), '*purple*' shaded violin plot appears to have the smallest spread while and the '*yellow*' shaded violin plot appears to have the longest spread. This implies that the Aggregation Scheme 4 estimate load values with error distributed at most in the ranges of 15 kW whereas for Aggregation V scheme some values are estimated with high error ranges. Box plot spreads are compact and comparatively similar for all the cases indicating most of the error (more than 75%) are within the ranges of 10kW. Also, the peak around 5 kW further suggests out of these 75% values most are distributed to an error around 5 kW, which is considerably low.

For Between 40-60 kW (Figure 5, Middle Panel), violin plots are mostly spread in the ranges of 20 kW and most of the error values are in the range of 10 kW (more than 75%). Same as above, kernel distribution plots are mostly peaking round error values of 5 kW. Kernel distribution plots for Aggregation 1, 2 and 3 have mostly similar structure while distribution plot for Aggregation 4 appears to have a slightly flatter peak and smooth tail. Aggregation 5 again seems to have the longest spread and several outlier values.

Between 60-85 kW (Figure 5, Bottom Panel), the distribution shape of violin plots has a bimodal shape for Aggregation 1, 3 and 5. Also tail for Aggregation 2 and 4 is smoothly decaying rather than a sharp decay notice above. Box plots are within the range of 20kW (suggesting around 75% values less than 20 kW error). Box plots appear to have the smallest spread for Aggregation 2. Aggregation scheme 5, once again is underperforming with the highest median value, the longest spread of error and two modes.

Nevertheless, in all the cases results are encouraging and proposed modelling scheme appears to simulate the dynamics of aggregated demand profiles with high accuracies in all the demand ranges. These results indicate that a simple mean/median-based or a simple time-of-use based features can be used to simulate a k-mean based clustering module for selecting a suitable sample for achieving optimum results with demand aggregation.

Conclusion

This paper intended to investigate the potentials for a k-mean based clustering approach to support demand synthesising tools/models and designing of an aggregation schematic for community-level energy demand modelling. Five different variants of the k-mean clustering are constructed using some basic/standard information available for 74 dwellings within a case-study community. A logical framework is designed to select a suitable sample from the clusters using some statistical and clustering information. The sample dwellings are simulated using a demand synthesising tool (HMM_GP). The demand synthesising tool (HMM_GP) is a purely data-driven system of statistical approaches that simply needed a continuous time series of electricity demand to simulate a user-specified number of synthetical demand series. The simulated series owns the same statistical characteristics as the parent series and thus represent a realistically possible scenario, which can be attributed to a different household with similar statistical properties. In this context, this paper provided a logical framework for identifying and selecting sample dwellings (that can optimally capture the diversity of the community) for demand aggregation module in the community demand modelling.

The paper presented a thorough examination of clustering results by comparing various performance indicators and graphs. All the schemes performed reasonably well, suggesting a wide range of information can be used for designing smart aggregation schemes if processed effectively with k-means. Corresponding to each of the clustering variants an aggregation schematic is constructed and further investigated. To assess the performance of aggregations scheme and role of clustering variants in providing a strong sample for demand aggregation, five synthetics aggregation demand series are generated and thoroughly analysed. All schemes appear to provide encouraging results thus confirming the potentials of a k-mean based clustering approach for constructing aggregation schematics. Most interestingly, all these findings collectively provide enough pieces of evidence to reject the hypothesis "*k-means clustering if involves a feature that accounting in time periods with a comparatively large volume of peak demand activities can more effectively predict peaks demand in aggregated profiles".*

Finally, the paper demonstrates the scope for further investigating a range of socio-economic and geomorphic information/factors in improving the potentials of data-driven modelling schemes for community-level demand modelling. Since a very basic analysis and clustering structure (using only two factors) is used, this work has

immense potential for future investigation with a different form of clustering approaches and clustering with several factors, such as clustering with PCA (Ding and He 2004).

Acknowledgement

This work is done as part of the EPSRC funded project - Community-scale Energy Demand Reduction in India (CEDRI: EP/R008655/1).

Bibliography

Bhattacharyya, S. C., and G. R. Timilsina. 2010. "A review of Energy System Models." *Journal of Energy Sector Management* 4 (4): 494-518.

Bhattacharyya, Subhes C., and Govinda R. Timilsina. 2009. "Energy Demand Models for Policy Formulation: A Comparative Study of Energy Demand Models." World Bank Policy Research working paper no. WPS 4866. Accessed 11 22, 2018. https://openknowledge.worldbank.org/bitstream/handle/10986/4061/WPS4866.pdf?sequence=1&isAllowed=y.

Debnath, K. B., D. P. Jenkins, S. Patidar, and A. Peacock. 2020. "Understanding Residential Occupant Cooling Behaviour through Electricity Consumption in Warm-Humid Climate." *Buildings* 10 (4): 78.

Ding, Chris, and Xiaofeng He. 2004. "K-means Clustering via Principal Component Analysis." *Proceedings of the 21st International Conference on Machine Learning.* Canada.

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

Hartigan, John A. 1975. *Clustering Algorithms.* 605 Third Ave. New York, NYUnited States: John Wiley & Sons, Inc.

Hintze, Jerry L., and Ray D. Nelson. 1998. "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician* 52 (2): 181-184.

Howell, Kayt. 2020. *Fintry Development Trust.* Accessed 08 21, 2020. http://fintrydt.org.uk/.

2017. *Industrial Strategy: building a Britain fit for the future.* HM Government, London: Government of UK.

Kassambara, Alboukadel. 2017. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning.* htp://www.sthda.com: STHDA.

Kshetri, Nir, and Jeffrey Voas. 2018. "Blockchain in Developing Countries." *IT Professional* (IEEE) 20 (2): 11-14.

Larose, Daniel T., and Chantal D. Larose. 2014. "Chapter 10: Hierarchical and k-means clustering ." In *Discovering Knowledge in Data: An Introduction to Data Mining*, by Daniel T. Larose and Chantal D. Larose, 209-227. New Jersey: IEEE computer Society, John Wiley & Sons.

McKenna, Eoghan, Sarah Higginson, Philipp Grunewald, and Sarah J. Darby. 2018. "Simulating residential demand response: Improving socio-technical assumptions in activity-based models of energy demand." *Energy Efficiency* (Springer Link) 11: 1583–1597.

Nisbet, Robert, Gary Miner, and Ken Yale. 2018. "Chapter 7 - Basic Algorithms for Data Mining: A Brief Overview." In *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*, 121-147. Academic press, Elsevier.

Patidar, Sandhya, David Paul Jenkins, Andrew Peacock, and Peter Mccallum. 2019. "Time Series Decomposition Approach for Simulating Electricity Demand Profile." *Building Simulation 2019, 16th IBPSA International International Conference.* Rome, Italy.

Pedersen, Thomas Lin. 2020. *Create Elegant Data Visualisations Using the Grammar of Graphics.* User manual, https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf: R - package "ggplot2".

Smith, Jackie. 2018. *http://smartfintry.org.uk/about-smart-fintry/resources/.* Smart Fintry Innovation Report. 13 04. Accessed 07 2020. http://smartfintry.org.uk/wp-content/uploads/2018/04/Smart-Fintry-Innovation-Report-final.pdf.

Stankovic, L., V. Stankovic, J. Liao, and C. Wilson. 2016. "Measuring the energy intensity of domestic activities from smart meter data." *Applied Energy* 183: 1565-1580.

Suganthi, L., and A. A. Samuel. 2012. "Energy models for demand forecasting - A review." *Renewable and Sustainable Energy Reviews* 16: 1223-1240.